Constructing Bayesian networks for criminal profiling from limited data

K. Baumgartner\textsuperscript{a,1}, S. Ferrari\textsuperscript{a,2}, G. Palermo\textsuperscript{b,3}

\textsuperscript{a}Pratt School of Engineering, Duke University, P.O. Box 90300, 176, Hudson Hall, Research Drive, Durham, NC 27708-0005, USA
\textsuperscript{b}Psychiatry and Neurology Department, Medical College of Wisconsin, Milwaukee, WI 53226, USA

\textbf{A R T I C L E I N F O}

Article history:
Received 21 March 2007
Accepted 21 March 2008
Available online xxxx

Keywords:
Criminal profiling
Crime analysis
Automation
Bayesian networks
Performance metrics

\textbf{A B S T R A C T}

The increased availability of information technologies has enabled law enforcement agencies to compile databases with detailed information about major felonies. Machine learning techniques can utilize these databases to produce decision-aid tools to support police investigations. This paper presents a methodology for obtaining a Bayesian network (BN) model of offender behavior from a database of cleared homicides. The BN can infer the characteristics of an unknown offender from the crime scene evidence, and help narrow the list of suspects in an unsolved homicide. Our research shows that 80% of offender characteristics are predicted correctly on average in new single-victim homicides, and when confidence levels are taken into account this accuracy increases to 95.6%.

\textcopyright 2008 Elsevier B.V. All rights reserved.

1. Introduction

The study of criminal behavior for the purpose of identifying the characteristics of an unknown offender and the motivation for the crime is commonly known as criminal profiling. In current practice, criminal profiling relies primarily on the personal experience of criminal investigators and forensic psychologists, rather than on empirical scientific methods [31]. As such, it may be subject to errors caused by cultural biases and misinterpretation [24,31,32,43]. After clearing a criminal case, investigators file the background characteristics and psychological diagnosis of the convicted offender together with the forensic evidence obtained from the crime scene. With the increased availability of computer and information technologies, law enforcement agencies have been able to compile databases with detailed offender and crime scene information from major felonies, such as murder, rape, and arson. Consequently, important authors have advocated that machine learning techniques will play a significant role in developing decision-aid tools for police investigations [4,17,27,32,42]. The most significant contributions to date have been recently reviewed in [17]. Rule-based systems have been proposed in [4] for knowledge acquisition from a database with modus operandi information. Research on inductive profiling has employed statistical analysis to classify offender behavior into categories or dichotomies, based on the crime scene evidence [12,25,34,35,37,39,41]. While this research has been successfully implemented to predict the approximate residence location of serial homicide offenders [35], it has been unable to identify psycho-behavioral offender profiles in single-victim (non-serial) homicides. This shortcoming has been attributed to the complexity of human behavior and to the large number of relevant variables, both of which limit the applicability of behavior classification techniques [2,31,32].

In this paper, a novel Bayesian network (BN) approach to criminal profiling is presented. The approach consists of learning a BN model of offender behavior from data and, subsequently, implementing the model for profiling by means of an inference engine. The database used in this paper is similar to the modus operandi database described in [4]. However, the BN approach is not limited by decisive “if-then” relationships, because it views the relationships among all variables as probabilistic. Unlike inductive profiling, the BN approach does not require to postulate behavior categories \textit{a priori} and, consequently, it is capable of identifying psycho-behavioral profiles in single-victim single-offender homicides (Section 6). Also, the inferred offender characteristics include confidence levels that represent their expected accuracy. Thus, when provided with a BN profile, the police can easily establish what are the reliable predictions in the investigated case.

Implementing BN models for inference has proven valuable in many applications, including medical diagnosis, economic forecasting, biological networks, and football predictions [1,19,20,23,30]. This literature shows that the effectiveness of BN inference and prediction is highly dependent on the sufficiency of the training database. While various approaches have been proposed for dealing with insufficient databases [11,14–16,21,26,30,40], there are no general guidelines for establishing whether a given database
is insufficient. In [45], it was shown that the size of a sufficient database depends on the number of variables, their domain, and the underlying probability distributions. But, while the variables and the domain definitions are known from the problem formulation, the underlying probability distribution is often unknown a priori. This paper presents a set of performance metrics that can be used to determine the sufficiency of an available database without knowledge of the underlying joint probability distributions (Section 4). Although a police database may include hundreds of cleared cases, they may still be insufficient to train a BN model due to the large number of relevant variables, and to the complexity of their relationships [3]. Therefore, in Section 5 these performance metrics are implemented to determine the size of a sufficient database with single-victim single-offender homicides. Subsequently, a BN model is trained using a newly modified K2 algorithm that improves performance once the database size is fixed (Section 5). In Section 6, the trained BN model is applied to infer the characteristics of unknown offenders from the crime scene evidence. The results show that when the confidence level is taken into account, the average accuracy of the BN predictions is 95.6%. For comparison, the evidence from two homicide cases has been presented to a team of expert criminologists. Based on the evidence alone, the experts predict 53% of all offender variables correctly. Whereas, in the same two cases the BN predicts 86% of all offender variables correctly, and displays 80% average accuracy in 1000 other homicide cases. Also, offender characteristics that cause disagreement among the experts are predicted correctly with a high confidence level by the BN. Finally, the structure of the BN model indicates what are the most significant relationships among the variables and, thus, it could be used for the scientific development of hypothesis on criminal psychology.

2. Background on Bayesian network inference and training

A Bayesian network (BN) approximates the joint probability distribution for a multivariate system based on expert knowledge and sampled observations that are assimilated through training [18,22]. A BN consists of a directed acyclic graph (DAG) and an attached parameter structure comprised of conditional probability tables (CPTs) that together specify a joint probability distribution [22]. The DAG \( A = (\mathcal{X}, \mathcal{F}) \) is composed of a set of directed arcs \( \mathcal{F} \) that represent the dependencies among a set of variables or nodes \( \mathcal{X} = \{X_1, \ldots, X_n\} \) known as universe, such that \( \mathcal{F} = \{(X_i, X_j) | X_i \in \mathcal{X}, X_j \neq X_i, j > i\} \). A node \( X_i \) represents an event, proposition, or mathematical quantity that has a finite number of mutually exclusive instantiations (denoted by lower case letters), and is said to be in its \( j \)th instantiation when \( X_i = x_{ij} \). The probability structure is that attached to \( A \), where \( \theta_i \) is the conditional probability \( p(X_i|\pi_i) \) attached to node \( X_i \) and the set \( \pi_i \) represents the immediate parents of \( X_i \).

In this research, the nodes \( \mathcal{X} \) and their instantiations are defined by criminologists and psychologists. The BN arcs \( \mathcal{F} \) and parameters \( \theta \) are learned from the database \( \mathcal{F} \) in this sequence. Structural training determines the set of arcs that "best" describes the database by considering all possible arcs between the nodes. The compatibility of each hypothesized structure with the training data is assessed by a scoring metric that approximates the conditional probability of \( \theta \) given \( \mathcal{F} \), \( p(\theta|\mathcal{F}) \) [18]. Since \( p(\mathcal{F}) \) is independent of \( \theta \), the joint probability \( p(\mathcal{F}, \theta) \) can be maximized in place of \( p(\mathcal{F}|\theta) \). A tractable scoring metric, known as K2, is obtained from \( p(\mathcal{F}, \theta) \) using the assumptions in [6], which include fixed ordering of variables in \( \mathcal{F} \):

\[
\mathcal{F} = \log \left( \prod_{i=1}^{n} \frac{(r_i - 1)!}{N_{ij} + r_i - 1)!} \prod_{j=1}^{N_{ij}} \right),
\]

where \( r_i \) is the number of possible instantiations of \( X_i \), and \( q_i \) is the number of unique instantiations of \( N_{ij} \), is the number of cases in \( \mathcal{F} \) in which \( X_i = x_{ij} \) and \( N_{ij} = \sum_{k=1}^{r_i} N_{ijk} \). Then, the BN structure that displays the highest compatibility with the data is sought by maximizing (1). Subsequently, the structure is held fixed, and the CPTs are computed by the Maximum Likelihood Estimation algorithm (MLE) [reviewed in [8,33]].

The BN \( (A, \theta) \) represents a factorization of the joint probability over a discrete sample space,

\[
p(\mathcal{F}) = p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i|\pi_i),
\]

for which all probabilities on the right-hand side are given by the CPTs. Therefore, when a variable \( X_i \) is unknown or hidden, Bayes’ rule of inference can be used to calculate the posterior probability distribution of \( X_i \) given evidence of the set of \( l \) variables, \( \mu_i \subset \mathcal{F} \), that are conditionally dependent on \( X_i \).

\[
p(X_i|\pi_i) = \frac{p(\pi_i|X_i)p(X_i)}{p(\pi_i)},
\]

where \( p(X_i) \) is the prior probability of \( X_i \). The likelihood function is factored as \( p(\pi_i|X_i) = \prod_j p(\pi_j|X_i) \), where \( \mu_j \) is the evidence of the \( j \)th variable in \( \mu_i \). The marginalization required to obtain \( p(\pi_i) \) is simplified using (2):

\[
p(\pi_i) = \sum_{k} p(\pi_i|X_k) = \sum_{k} p(X_k = x_{ik}) \prod_j p(\pi_j|X_i),
\]

The posterior probability in (3) is used to obtain the prediction \( \mathcal{X} = \arg \max_X p(X = x_{ik}|\pi_i) \), and its posterior probability is the confidence level of the prediction. Furthermore, by identifying conditional independencies among nodes from the so-called Markov separation properties, inference of hidden variables can be completed efficiently even in large networks [9].

Bayesian networks are particularly well suited to criminal profiling because they learn from data, and utilize the experience of criminologists in selecting the nodes and node ordering. The confidence levels provided for the offender profile inform detectives of the likely accuracy of each prediction. In addition, the graphical structure of the BN represents the most significant relationships between offender behavior and crime scene actions, which may be useful in developing new scientific hypothesis on criminal behavior.

3. Bayesian network approach to criminal profiling

This research develops an approach for obtaining a BN model of criminal behavior that (1) captures the most significant relationships among the relevant criminal profiling variables, and (2) is used to predict the profile of an unknown offender given evidence from the crime scene. The methodology consists of using expert knowledge to define the BN universe, and the fixed node ordering for structural training (as shown in Section 2). The universe \( \mathcal{F} \) consists of 57 binary variables that have been identified as relevant to the criminal process by criminal investigators and forensic psychologists. A sample of these variables is illustrated in Table 1, and the complete list is shown in [38]. Each variable \( X_i \in \mathcal{F} \) is binary, and represents a characteristic or event that is either present or absent at the crime scene. \( \mathcal{F} \) is partitioned into set \( \mathcal{E} = \{E_1, \ldots, E_k\} \) containing \( k = 36 \) evidence variables that are observable from the crime scene, and set \( \mathcal{Y} = \{Y_1, \ldots, Y_m\} \) containing \( m = 21 \) offender variables that characterize the offender and, thus, are unknown or hidden at the crime scene.

The BN model structure, \( \mathcal{F} \), and parameters, \( \theta \), are learned using a police database of cleared single-victim single-offender homicides, \( \mathcal{D} = \{C_1, \ldots, C_k\} \). Each case \( C_i \) is a complete observation

Please cite this article in press as: K. Baumgartner et al., Constructing Bayesian networks for criminal profiling from limited data, Knowl. Based Syst. (2008), doi:10.1016/j.knosys.2008.03.019
of $x$ that is obtained from a cleared homicide case. The values of all variables in $\mathcal{Y}$ are obtained from interviews with the convicted offender, and the values of all variables in $\mathcal{E}$ are obtained from the police record of the investigation. Then, $\mathcal{E}$ is randomly partitioned into a training set $\mathcal{E}$ and a validation set $\mathcal{V}$, such that $\mathcal{E} \cap \mathcal{V} = \emptyset$. Thus, none of the validation cases are used for training and can be considered to be new to the BN model.

A set of performance metrics is developed in the next section to establish the size of $\mathcal{E}$ that is sufficient to train the BN model by means of a modified K2 algorithm presented in Section 5. Finally, the BN model is used to produce psycho-behavioral offender profiles in new criminal cases (Section 6). In this research, the new cases are taken from $\mathcal{V}$ in order to determine the accuracy of the BN predictions. However, since $\mathcal{E} \cap \mathcal{V} = \emptyset$, the same approach and accuracy would apply to new unsolved cases, such as those encountered by investigators.

Let $\mathcal{E} = \{e_1, \ldots, e_k\}$ denote the evidence obtained from the crime scene of a new case, where $\mathcal{E}$ is the observed value of $E_i$ (see Table 1 for examples). Then, the junction tree inference engine [10], implemented by the MATLAB Bayesian Network Toolbox function *tree_inf_engine* [29], is used to compute the posterior probability distribution for every offender variable $Y_i \in \mathcal{Y}$. An offender variable prediction, $\hat{Y}_i$, is the instantiation with the largest posterior probability, $\hat{Y}_i = y'_i \equiv \arg \max_{y'_i} p(Y_i = y'_i | \mathcal{E})$, and $p(y'_i | \mathcal{E})$ is said to be the confidence level (CL) of the prediction $\hat{Y}_i$. The set of all predictions, $\mathcal{\hat{Y}} = \{\hat{Y}_1, \ldots, \hat{Y}_{2^n}\}$, and corresponding confidence levels is referred to as *BN profile*. The results in Section 6 indicate that this BN approach is able to produce accurate psycho-behavioral profiles in single-victim single-offender homicides and, thus, is a promising decision-support tool for police investigations.

### 4. Development of Bayesian network performance metrics

The performance of Bayesian network training algorithms always depends on the sufficiency of the training database. A sufficient database is representative of the statistical population and sample complexity [5]. Since the type of criminal cases that are solved and recorded in a police database cannot be controlled, we are interested in determining whether the database size is sufficient. The size of a criminal profiling database can be increased by obtaining data from different law enforcement agencies. However, due to the legalities associated with criminal records and to the non-uniformity of the agencies protocols, obtaining new data is both difficult and expensive. In [45], it was shown that the size of a sufficient BN database depends on the number of nodes, the size of their domain, and the underlying probability distributions. While the nodes and their domain are known from the problem formulation, the underlying probability distribution is typically unknown a priori. Therefore, in this section, we present a set of metrics that can be used to determine whether an available database is sufficient without obtaining new data, and without knowledge of the underlying probability distributions.

An important use of a BN model of offender behavior is the prediction of a criminal profile in a new, unsolved homicide. We define the *predictive accuracy* of a BN to be the average frequency at which the hidden variables $\mathcal{Y}$ are predicted correctly from the evidence $\mathcal{E}$, when their predicted value is equal to the instantiation with maximum posterior probability. A common approach to viewing the performance of the training algorithm is to plot the prediction performance versus the size of the training set, obtaining a so-called *learning curve* [36]. If a BN accurately represents the joint probability distribution over $x$, then its predictive accuracy for $m$ hidden variables is,

$$Q_{\text{BN}}(\mathcal{Y}) = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n_y} \max_{y'_j} p(y'_j | \mathcal{E}_i),$$

where $\mathcal{E}_i$ denotes the simultaneous occurrence of a set of instantiations of $\mathcal{E}$, and thus $j = 1, \ldots, n_y$, with $N_y = \prod_{i} n_i$. The proof and a simple example are provided in Appendix A. As the size of $\mathcal{E}$ becomes sufficient and the trained BN approaches the actual joint probability distribution, this learning curve approaches $Q_{\text{BN}}$.

Also, a lower bound of (5) that is independent of the underlying distribution is derived and used to establish whether the BN can be improved by additional training data. We define the *frequency of occurrence prediction* of a variable $Y_i$ to be the instantiation that occurs most frequently in the database $\mathcal{Y}$, and denote it by $Y^*_i$. If $Y_i$ is independent of any other variable in $\mathcal{Y}$, then $Q_{\text{BN}}$ is the optimal prediction, and the average accuracy for $m$ hidden variables is

$$Q_{\text{DF}}(\mathcal{Y} | \mathcal{E}) = \frac{1}{m} \sum_{i=1}^{m} \max_{y'_j} f(y'_j | \mathcal{E}_i),$$

where the frequency $f(y'_j | \mathcal{E}_i)$ is the number of cases in $\mathcal{Y}$ in which $Y_i = y'_j$, divided by $d$, and $\mathcal{E}_i$ denotes evidence of the instantiation. It is shown in Appendix B that if the variables in $\mathcal{Y}$ are not independent and the BN accurately represents the joint probability distribution over $x$, then $Q_{\text{DF}} \geq Q_{\text{DF}}$. It follows that when a trained BN displays $Q_{\text{DF}} < Q_{\text{DF}}$, it cannot represent the underlying probability distribution and, thus, the database $\mathcal{E}$ is insufficient. Therefore, $Q_{\text{DF}} \geq Q_{\text{DF}}$ is a necessary but not sufficient condition for deeming a database $\mathcal{E}$ sufficient for training.

When $Q_{\text{DF}} \geq Q_{\text{DF}}$, the structural robustness of the trained BN model can provide additional insight into the sufficiency of $\mathcal{E}$. A BN defined over a fixed universe $\mathcal{X}$ is said to be *structurally robust* if its arc structure is insensitive to small changes in the training set $\mathcal{E}$. Structural sensitivity analysis was first proposed in [45], where the error between the learned structure and the true structure has been shown to decrease with the size of the training database. We present a structural robustness metric that is independent of the true structure, since the true structure is typically unavailable when the underlying probability distribution is unknown. We represent the learned graphical structure of a BN with $n$ nodes in matrix form,

$$S = \begin{bmatrix} S_{1,1} & \ldots & S_{1,c} & \ldots & S_{1,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ S_{n,1} & \ldots & S_{n,c} & \ldots & S_{n,n} \end{bmatrix}$$

where each entry $s_{i,c}$ represents the presence ($\pm 1$) or absence (0) of an arc between two variables $X_i, X_c \in \mathcal{X}$. The direction of the arc is positive according to the expert node ordering and, thus, $s_{i,c} = +1$ and, $s_{c,i} = -1$ when $i < c$ and there exist an arc $X_i \rightarrow X_c$. Let $\mathcal{S}$ represent an available training set, and $\mathcal{S}$ represent $\mathcal{S}$ with 10% of cases randomly removed. The resulting number of arc differences is a pairwise comparison of the structures obtained from the two training sets, denoted by $\mathcal{S}$ and $\mathcal{S}$, respectively, such that:

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Definition of selected offender and crime scene variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>Definition</td>
</tr>
<tr>
<td>$Y_4$</td>
<td>Prior record of property damage</td>
</tr>
<tr>
<td>$Y_5$</td>
<td>Prior record of disorderly conduct</td>
</tr>
<tr>
<td>$Y_6$</td>
<td>Previous imprisonment or youth detention</td>
</tr>
<tr>
<td>$Y_7$</td>
<td>History of sex crime</td>
</tr>
<tr>
<td>$Y_{10}$</td>
<td>Record of armed services</td>
</tr>
<tr>
<td>$E_{11}$</td>
<td>Victim sustained stabbing wounds</td>
</tr>
<tr>
<td>$E_{12}$</td>
<td>Blunt instrument used on victim</td>
</tr>
<tr>
<td>$E_{13}$</td>
<td>Offender used own body as weapon (e.g. strangulation)</td>
</tr>
<tr>
<td>$E_{14}$</td>
<td>Victim was shot</td>
</tr>
<tr>
<td>$E_{15}$</td>
<td>Victim sustained wounds to head (excluding face and neck)</td>
</tr>
<tr>
<td>$E_{16}$</td>
<td>Victim sustained wounds to face (ears forward)</td>
</tr>
<tr>
<td>$E_{17}$</td>
<td>Victim was sexually assaulted</td>
</tr>
<tr>
<td>$E_{18}$</td>
<td>Arson to crime scene or body</td>
</tr>
<tr>
<td>$E_{19}$</td>
<td>Body was found in water</td>
</tr>
</tbody>
</table>

Please cite this article in press as: K. Baumgartner et al., Constructing Bayesian networks for criminal profiling from limited data, Knowl. Based Syst. (2008), doi:10.1016/j.knosys.2008.03.019
Thus, the metric $1/\Delta A_{10}$ can be used to represent BN structural robustness, and is expected to increase as the size of $\mathcal{F}$ increases. A lack of structural robustness typically indicates that $\mathcal{F}$ is not reliable, and that the parameters $\theta$ learned subsequently to $\mathcal{F}$ may also be inaccurate. Therefore, when the size of $\mathcal{F}$ becomes sufficient, $1/\Delta A_{10}$ is small and approximately constant, as demonstrated by the numerical results in Section 5.

Information theory has been applied to BNs for the purpose of improving the training process and for performance analysis [14,28,44,45]. For instance, the cross-entropy measure between a trained BN and a known underlying probability distribution is utilized in [45] for evaluating and comparing the learning performance of several training algorithms. In [14], conditional mutual information is used as a scoring function in a new structural training algorithm. Here, we present a different BN application of mutual information to obtain a metric quantifying the reduction in uncertainty in the hidden variables $\psi$ brought about by evidence of $\phi$. The mutual information between two sets of random variables, $I(\psi; \phi)$, represents the amount of information that $\phi$ has about $\psi$, and is related to the entropy as follows [7]:

$$I(\psi; \phi) = H(\psi) - H(\psi | \phi) = H(\psi) - H(\psi | \phi) = H(\psi) - H(\psi | \phi).$$

(9)

Thus, we let $I(\psi; \phi)$ denote the mutual information of a Bayesian network with hidden variables $\psi$ to be inferred from evidence of the variables $\phi$. The joint entropy of the BN can be computed using the BN factorization [2]:

$$H(\psi, \phi) = H(\psi) = H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | \pi_i).$$

(10)

Then, assuming that the evidence variables are conditionally independent (Section 5), a formula is derived expressing the BN mutual information in terms of the learned CPTs,

$$I(\psi; \phi) = \sum_{j=1}^{k} [H(E_j) - H(E_j | \pi_j)] + \sum_{i=1}^{m} [H(Y_i) - H(Y_i | \pi_i)]$$

$$= \sum_{j=1}^{k} \left\{ \sum_{i=1}^{n} p(\pi_j) p(e_j | \pi_j) \log[p(\pi_j) p(e_j | \pi_j)] \right\}$$

$$- p(\pi_j) p(e_j | \pi_j) \log[p(\pi_j) p(e_j | \pi_j)]$$

$$+ \sum_{i=1}^{m} \left\{ \sum_{i=1}^{n} p(\pi_i) p(y_i | \pi_i) \log[p(\pi_i) p(y_i | \pi_i)] \right\}$$

$$- p(\pi_i) p(y_i | \pi_i) \log[p(\pi_i) p(y_i | \pi_i)],$$

(11)

where the summations over $\pi_i, \pi_j \in \mathcal{F}$ denote marginalization. As can be seen from (11), $I(\psi; \phi)$ can be computed from the BN CPTs and, based on its definition, constitutes a performance metric for BN inference. Consequently, the mutual information learning curve will approach a constant value as the size of the database becomes sufficient.

When only one database $\mathcal{F}$ is available, several databases of increasing size can be obtained by sampling $\mathcal{F}$. These databases are used to plot learning curves for the performance metrics derived in this section. Then, these learning curves can be used in combination with $Q_m$ to determine whether $\mathcal{F}$ is sufficient, as shown in the next section.

5. Numerical studies and implementation of performance metrics

The performance metrics presented in the previous section are used to determine whether a database $\mathcal{F}$ with 5000 cases is sufficient for obtaining the BN model of criminal behavior described in Section 3. The universe $\phi$ and its partition into offender and evidence variables are described in Section 3. Databases of size $t$ varying from 50 to 5000 in increments of 50 cases are generated via sampling. For every database, a BN model is obtained through structural and parameter training algorithms. Then, the metrics $Q_m, 1/\Delta A_{10}$, and $I(\psi; \phi)$, are computed for each BN model and plotted on a learning curve. Structural training is performed using the K2 algorithm (Section 2) as well as a newly modified version, referred to as K2, which has shown to improve the performance of the criminal profiling BNs.

Fig. 1. A simple BN structure learned by the K2 algorithm (a) omitting arcs between evidence variables is compared to a K2 structure that is Markov equivalent (b), and to one that is not (c).

Please cite this article in press as: K. Baumgartner et al., Constructing Bayesian networks for criminal profiling from limited data, Knowl. Based Syst. (2008), doi:10.1016/j.knosys.2008.03.019
with the K2’ algorithm, \( Q_{BN_{K2'}} < Q_{FO} \) when the training database contains less than 250 training cases. Thus, any database with \( t < 250 \) is insufficient, based on \( Q_{FO} \) alone. Fig. 2 also shows the predictive accuracy learning curve for the K2’ algorithm. Since the K2 algorithm utilizes the training data less efficiently than the K2’, any database with \( t < 600 \) is insufficient. Also, the learning curve in Fig. 2 illustrates that both \( Q_{BN_{K2'}} \) and \( Q_{BN_{K2}} \) become approximately constant when \( t \approx 800 \). Therefore, based on the predictive accuracy metric, a sufficient database must contain at least 800 cases.

Structural robustness is evaluated by retraining each BN model after 10% of the cases has been removed from each training set sampled from \( \mathcal{F} \). Then, the structural robustness metric \( \Delta A_{10} \) in (8) is plotted on the learning curve in Fig. 3, where the abscissa represents the size \( t \) before 0.01\( t \) cases are removed. As in Fig. 2, the training sets sampled from \( \mathcal{F} \) have a size \( t \) that varies between 50 and 5000 in increments of 50. When the training set contains more than 1000 cases, \( 0 \leq \Delta A_{10} < 10 \) and becomes approximately constant. Also, these curves indicate that the robustness of BNs learned by the K2 algorithm is only slightly worse than that of BNs learned by the K2’. It can be seen from Fig. 3 that for both algorithms a sufficient database contains at least 1000 cases.

The BN mutual information (11) is used to obtain the learning curves in Fig. 4. The mutual information decreases as the amount
of training data increases because, when the data is insufficient, spurious arcs are introduced by the structural training algorithm. As the amount of data increases, these spurious dependencies are eliminated and the BN structure becomes more reliable and robust (Fig. 3). Therefore, when the underlying probability distributions are unknown the sufficiency of the training set can be established based on the sensitivity of the BN robustness and mutual information to the number of training cases. It can be seen from Fig. 4 that when the training set contains more than 500 cases, the BN mutual information becomes approximately constant. When \( t > 1000 \), the BN structure is robust, all performance metrics are approximately constant (Figs. 2–4), and \( Q_{BN} > Q_{FO} \). Therefore, it can be concluded that a database of size \( t_s = 1000 \) is sufficient for training a BN model of criminal behavior (as described in Section 3). Subsequently, this BN can be utilized to support criminal investigations as shown in the following section.

6. Bayesian network profiler for decision-support in criminal investigations

A trained BN model of offender behavior can constitute a valuable decision-support tool for police investigations. As schematized in Fig. 5, the evidence obtained from the crime scene of a new case is inserted in the trained BN model and, through the inference engine, the offender psycho-behavioral profile is produced. The BN profile consists of a set of predictions comprising the most likely values of the offender variables and the corresponding confidence levels (CLs). This information is then displayed to investigators on a computer screen or palm pilot to help narrow the list of suspects in an unsolved case, and identify the motive for the crime. As additional cases are cleared by the police, the BN can be updated through incremental training off line (dashed lines).

As an example, we analyze two homicide cases that are taken from \( V \) and, thus, have never been used for training the BN model. In the first case, the actual offender is a male with no prior criminal record, who had no prior relationship with the victim but is familiar with the crime scene. Table 2 shows a sample of the 21 offender variables comprising the profile \( \Phi \). Actual evidence of the 36 crime scene variables (not shown for confidentiality reasons) is provided to both the BN model and a team of experts in forensic psychiatry. The results of the BN and experts’ profiles are shown in Table 2 for the same sample variables. A “yes/no” answer indicates that “yes” is considered a more likely value for the variable, but there is dis...

Fig. 4. Mutual information learning curve for BNs obtained by the \( K_2 \) and \( K_2 \) algorithms, with \( \Delta t = 50 \).

Fig. 5. Implementation of BN model as decision-support tool for police investigations.
agreement among the experts. Whereas, the BN predictions are accompanied by a percent CL indicated in parenthesis. For the sample variables in Table 2, the BN produces only one incorrect prediction, \( Y_8 \), compared to three incorrect predictions by the experts. Even more importantly, \( Y_8 \) carries a very low confidence, CL = 51%, indicating to a potential user that there is a fairly high probability (0.49) that this prediction is incorrect. Also, when all 21 offender variables are considered, the BN displays a much higher percent accuracy (90.5%) than the team of experts (62%).

In a second homicide case illustrated in Table 3 the actual offender is a female with no prior criminal record, who is familiar with the crime scene, and is a blood relative of the victim. This is considered to be a more difficult case, partly because the offender is a female while the gender of single-victim homicide offenders is predominately male (e.g. 91% of cases in \( \phi \)). Although both the BN and the team of experts predict the offender gender correctly, they also produce more incorrect predictions than in the first case (Table 3). It can be seen that the BN is not only able to correctly identify the gender, but also provides a high confidence in this prediction, CL = 99%. When all 21 offender variables are considered, the BN displays a much higher average percent accuracy (85.7%) than the team of experts (47.6%), and its predictions carry high confidence levels on average (86.2%). It was found that in these two homicide cases, the BN model displayed a high average confidence level, CL = 82.6%, for the variables that are predicted incorrectly by the experts.

The BN model was tested using the entire validation set \( \psi \) containing 1,000 single-victim homicide cases and each reporting 21 offender variables. The results show that 80% of offender characteristics are predicted correctly on average. Moreover, since the accuracy of the prediction increases with the confidence level, higher accuracy can be obtained by considering only those predictions with a high confidence level. This result is shown in Table 4, where the average accuracy is shown for offender variable predictions that are organized by confidence level. Also, the accuracy range obtained for variables that, on average, display a particular confidence level range is shown in Table 5. It can be concluded that the confidence levels are representative of the accuracy of individual predictions, as well as of offender variables, on average. In other words, if in an unsolved case an offender variable is predicted with \( CL \geq 70 \), then its accuracy is approximately 85.9% (Table 4).

Also, if an offender variable, such as \( Y_{11} \), displays a confidence level \( 80 \% \leq CL < 90 \% \), on average, then its accuracy varies between 79.4% and 92.4% in all cases considered in this study (Table 5). Finally, the average percent accuracy of a sample of offender variables is shown in Table 6. These results illustrate what offender characteristics are typically predicted incorrectly by the BN. This could be due to the data not being representative of their relationships with the other variables in the universe \( \phi \), for example due to bias and collection errors, or to these relationships being weak. Therefore, these offender variables can be the subject of future research. On the other hand, characteristics such as the age and gender of the offender and family relationship with the victim are typically predicted correctly by the BN and, therefore, can be used to narrow the list of suspects in unsolved cases.

Another benefit of BN modeling is the graphical display of the relationships learned from data. A slice of the trained BN model structure, which represents the most significant relationships between the criminal profiling variables, is shown in Fig. 6. For example, the arc between \( Y_{13} \) and \( E_{27} \) indicates that hiding the victim’s body outdoors is influenced by the offender being ac-

---

Table 2: BN profile and profile produced by a team of forensic psychiatry experts for a single-victim homicide

<table>
<thead>
<tr>
<th>Variables</th>
<th>Actual</th>
<th>BN (CL %)</th>
<th>Expert team</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y_1 ) = prior record of violence</td>
<td>No</td>
<td>No (74)</td>
<td>No</td>
</tr>
<tr>
<td>( Y_2 ) = prior record of damage</td>
<td>No</td>
<td>No (84)</td>
<td>No</td>
</tr>
<tr>
<td>( Y_3 ) = unemployed</td>
<td>Yes</td>
<td>No (51)</td>
<td>Yes</td>
</tr>
<tr>
<td>( Y_{11} ) = familiar with CS</td>
<td>Yes</td>
<td>Yes (84)</td>
<td>Yes</td>
</tr>
<tr>
<td>( Y_{12} ) = gender</td>
<td>Male</td>
<td>Male (73)</td>
<td>Male</td>
</tr>
<tr>
<td>( Y_{14} ) = prior record of abuse</td>
<td>No</td>
<td>No (80)</td>
<td>Yes</td>
</tr>
<tr>
<td>( Y_{17} ) = prior record of fraud</td>
<td>No</td>
<td>No (67)</td>
<td>No</td>
</tr>
<tr>
<td>( Y_{19} ) = prior sexual relationship with victim</td>
<td>No</td>
<td>No (57)</td>
<td>Yes</td>
</tr>
<tr>
<td>( Y_{20} ) = blood relative of victim</td>
<td>No</td>
<td>No (90)</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Average accuracy (CL %)</td>
<td>–</td>
<td>90.5% (79)</td>
<td>62%</td>
</tr>
</tbody>
</table>

Table 3: BN profile and profile produced by a team of forensic psychiatry experts for a single-victim homicide

<table>
<thead>
<tr>
<th>Variables</th>
<th>Actual</th>
<th>BN (CL %)</th>
<th>Expert team</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y_1 ) = prior record of violence</td>
<td>No</td>
<td>No (83)</td>
<td>Yes/No</td>
</tr>
<tr>
<td>( Y_2 ) = prior record of damage</td>
<td>No</td>
<td>No (95)</td>
<td>Yes/No</td>
</tr>
<tr>
<td>( Y_3 ) = unemployed</td>
<td>Yes</td>
<td>Yes (66)</td>
<td>Yes/No</td>
</tr>
<tr>
<td>( Y_{11} ) = familiar with CS</td>
<td>Yes</td>
<td>Yes (88)</td>
<td>Yes</td>
</tr>
<tr>
<td>( Y_{12} ) = gender</td>
<td>Female</td>
<td>Female (99)</td>
<td>Female</td>
</tr>
<tr>
<td>( Y_{14} ) = prior record of abuse</td>
<td>No</td>
<td>No (80)</td>
<td>Yes</td>
</tr>
<tr>
<td>( Y_{17} ) = prior record of fraud</td>
<td>No</td>
<td>No (80)</td>
<td>Yes</td>
</tr>
<tr>
<td>( Y_{19} ) = prior sexual relationship with victim</td>
<td>No</td>
<td>No (90)</td>
<td>Yes</td>
</tr>
<tr>
<td>( Y_{20} ) = blood relative of victim</td>
<td>Yes</td>
<td>Yes (66)</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Average accuracy (CL %)</td>
<td>–</td>
<td>85.7% (86)</td>
<td>47.6%</td>
</tr>
</tbody>
</table>

Table 4: Average BN predictive accuracy as a function of confidence level

<table>
<thead>
<tr>
<th>Confidence level (%)</th>
<th>Correct</th>
<th>Total (number of predictions)</th>
<th>Average accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( CL \geq 50 )</td>
<td>16,511</td>
<td>21,000</td>
<td>78.6</td>
</tr>
<tr>
<td>( CL \geq 60 )</td>
<td>15,054</td>
<td>18,361</td>
<td>82</td>
</tr>
<tr>
<td>( CL \geq 70 )</td>
<td>12,845</td>
<td>14,952</td>
<td>85.9</td>
</tr>
<tr>
<td>( CL \geq 80 )</td>
<td>10,515</td>
<td>11,802</td>
<td>89.1</td>
</tr>
<tr>
<td>( CL \geq 90 )</td>
<td>5,084</td>
<td>5,321</td>
<td>95.6</td>
</tr>
</tbody>
</table>

Table 5: Accuracy range for offender variables that display, on average, a confidence level within the specified ranges

<table>
<thead>
<tr>
<th>Confidence level range (%)</th>
<th>Accuracy range (%)</th>
<th>Number of offender variables (Example)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50–60</td>
<td>55.8</td>
<td>1, ( Y_6 ) = unemployed</td>
</tr>
<tr>
<td>60–70</td>
<td>63.8–67.8</td>
<td>4, ( Y_{11} = \text{prior sexual relationship} )</td>
</tr>
<tr>
<td>70–80</td>
<td>71.0–77.7</td>
<td>5, ( Y_6 = \text{prior record of burglary} )</td>
</tr>
<tr>
<td>80–90</td>
<td>79.4–92.4</td>
<td>8, ( Y_{13} = \text{familiarity with crime scene} )</td>
</tr>
<tr>
<td>90–100</td>
<td>97.0–98.8</td>
<td>3, ( Y_{20} = \text{blood relative of victim} )</td>
</tr>
</tbody>
</table>

Table 6: Average percent accuracy of a selected group of offender variables

<table>
<thead>
<tr>
<th>Offender variable</th>
<th>Average percent accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y_{10} ) = sexual relationship with victim</td>
<td>63.8%</td>
</tr>
<tr>
<td>( Y_{11} ) = prior record of fraud</td>
<td>67.7%</td>
</tr>
<tr>
<td>( Y_6 ) = history of psychiatric/social problems</td>
<td>67.8%</td>
</tr>
<tr>
<td>( F_{11} ) = prior record of fraud</td>
<td>71.0%</td>
</tr>
<tr>
<td>( Y_5 ) = prior record of violence</td>
<td>73.6%</td>
</tr>
<tr>
<td>( Y_4 ) = prior record of burglary</td>
<td>77.6%</td>
</tr>
<tr>
<td>( Y_3 ) = prior record of damage</td>
<td>77.7%</td>
</tr>
<tr>
<td>( Y_{14} ) = prior record of abuse</td>
<td>79.4%</td>
</tr>
<tr>
<td>( Y_{16} ) = young offender, 17–21 years old</td>
<td>82.8%</td>
</tr>
<tr>
<td>( Y_{15} ) = familiarity with crime scene</td>
<td>86.1%</td>
</tr>
<tr>
<td>( Y_{12} ) = male gender</td>
<td>89.7%</td>
</tr>
<tr>
<td>( Y_{20} ) = blood relative to victim</td>
<td>91.1%</td>
</tr>
</tbody>
</table>
quainted with the victim prior to the crime. As another example, the arc between \(Y_{12}\) and \(E_{23}\) indicates that the gender of the offender influences whether the weapon is brought to the crime scene. This finding is consistent with our understanding of female offenders, who typically employ weapons that are already present at the scene, such as knives from the kitchen. The relationship between the existence of a prior sexual relationship between the offender and the victim \((Y_{19})\) and the victim being blindfolded during the crime \((E_{10})\) is also known to forensic psychiatrists. This is attributed to the offender wanting to avoid eye contact and feelings of shame brought about by his or her familiarity with the victim. On the other hand, the influence that the offender’s prior record of fraud \((Y_{17})\) and employment status \((Y_{18})\) have on the presence of victim suffocation \((E_{32})\) is surprising to the experts. A possible explanation is that such an offender may feel inferior due to unemployment and find manual strangulation a source of empowerment.

Through the inference engine, the influence of the offender characteristics on the behavior exhibited at the crime scene and reflected in the evidence variables is taken into account automatically. But, another important advantage of this approach is that the trained BN structure can be easily interpreted and utilized by forensic psychiatrists for conducting research on the psychological mechanisms underlying criminal behavior.

7. Conclusions

The increased availability of computer and information technologies has enabled law enforcement agencies to compile extensive databases with detailed information about major felonies, such as murder, rape, and arson. Consequently, several authors have advocated that machine learning techniques will play a significant role in developing decision-aid tools for police investigations [4, 17, 27, 32, 42]. The most significant contributions to date have been recently reviewed in [17]. In this paper, we develop an approach for obtaining BN decision-aid tools that consists of the following steps: (1) assessing the sufficiency of an available database; (2) training a BN model using both expert knowledge and data; and, (3) implementing an inference engine to produce offender profiles in unsolved cases. Numerical studies demonstrate that the BN model can be used to successfully infer the characteristics of an unknown offender from the crime scene evidence in single-victim homicides. On average, 80% of the offender characteristics are predicted correctly by the BN profile. Moreover, since each prediction is accompanied by a confidence level that is proportional to its expected accuracy, by considering only predictions with high confidence levels, the average accuracy increases to 95.6%. Hence, the BN profile can be implemented by investigators to narrow the list of suspects in unsolved homicides, and identify the motivation for the crime.

Acknowledgments

The authors wish to thank Antony Pinizzotto, Gabrielle Salfati, Marco Strano, and Roberta Bruzzone for their invaluable suggestions and contributions to this study, which was conducted in collaboration with the International Crime Analysis Association.

Appendix A: Bayesian network predictive Accuracy

First, consider the case of one hidden variable, \(Y_e\), and let \(\varpi = \{Y_j/Y_i \neq Y_j \neq i\}\). The prediction \(\hat{Y}_i\) is obtained by inferring \(Y_i\) from evidence \(e\) such that,

\[
\hat{Y}_i = y'_i = \arg \max p(y_i'|e) = \arg \max p(y_i,e), \quad i = 1, \ldots, r_i. \tag{A.1}
\]

Therefore, the predictive accuracy of one hidden variable is,

\[
Q_{BN}(Y_i) = p(y_i'|e) p(e). \tag{A.2}
\]

and \(Q_{BN} \in [0,1]\) for any \(r_i\). Since there are \(N_e = \prod_{i=1}^{r_i} n_e\) possible evidence combinations, the probability of any simultaneous occurrence of a set of instantiations \(\vec{y} = (e_{11} \cup \ldots \cup e_{k_e})\) is the prior \(p(\vec{y})\), where \(j = 1, \ldots, N_e\). Then, the predictive accuracy of \(Y_i\) can also be written in terms of the BN CPTs, as follows,

\[
Q_{BN}(Y_i) = \sum_{j=1}^{N_e} \max \{p(y_i'|\vec{y}_j) p(\vec{y}_j) = \sum_{j=1}^{N_e} \max \{p(\vec{y}_j|y_i) p(y_i)\}. \tag{A.3}
\]

where \(p(\vec{y}_j) = \sum_{i=1}^{r_i} p(n_i)p(y_{i1}|n_i)\). Then, for a set of \(m\) hidden variables, \(\varpi = \{Y_1, \ldots, Y_m\}\), the BN predictive accuracy is the average of the individual predictive accuracies.
As a simple example, consider the two-node BN shown in Fig. 7. Y is the hidden variable to be inferred from E, and both variables are binary. Then, the predictive accuracy of Y is computed from (5) as follows,

\[
Q_{\text{BN}}(Y) = \frac{1}{m} \sum_{i=1}^{m} \max_{\gamma_i} p(y_i | \bar{y}_i) = \frac{1}{m} \sum_{i=1}^{m} \max_{\gamma_i} \left( p(\gamma_i | y_i) \sum_{n=1}^{N_y} p(y_i, \bar{n}_i | n_i) p(n_i) \right).
\]

This gives the same result as (A.4).

Appendix B. Frequency of occurrence lower bound

Consider one hidden variable Y, and a set of k evidence variables $\{e\}$. Then, the predictive accuracy of a prediction $\bar{Y}$ is given by (A.3), where $\{e\} = \{e_1, \ldots, e_r\}$ and $N_e$ is the number of all possible evidence combinations. Similarly to (A.1), based on the definition of prediction we let $y^{(0)}$ denote the instantiation chosen as the highest posterior probability when $\{e\} = \gamma_i$, namely:

\[
y^{(0)} = \arg \max_{\gamma_i} \{p(y_i | \bar{y}_i)\} \quad \text{arg max} \{p(y_i | \bar{y}_i)\}
\]

using Bayes' rule. It follows that $p(y^{(0)} | \bar{y}^{(0)}) = p(\gamma_i | y_i) p(y_i)$ for all $i = 1, \ldots, r$. Also (A.3) can be written as:

\[
Q_{\text{BN}}(Y) = \sum_{i=1}^{m} \max_{\gamma_i} \left( p(\gamma_i | y_i) \right).
\]

Now let $y^{(0)}$ denote the instantiation chosen as the frequency of occurrence prediction $\bar{Y}$, such that $y^{(0)} = \arg \max \{p(y_i | \bar{y}_i)\}$, where both instantiations $y^{(0)}$ and $y^{(0)}$ belong to the domain of $Y$, which contains $r$ mutually exclusive instantiations denoted by $\gamma_i$ (Section 2). Then, if $y^{(0)} = \bar{y}^{(0)} \forall i = 1, \ldots, N_e$ and the database $\bar{Y}$ is representative of the statistical population, the predictive accuracy (B.2) simplifies to:

\[
Q_{\text{BN}}(Y) = \sum_{i=1}^{m} \max_{\gamma_i} \left( p(\gamma_i | y_i) \right) = \sum_{i=1}^{m} \max_{\gamma_i} \left( p(\gamma_i | y_i) \right).
\]

achieving the lower bound $Q_{\text{BN}}$ in (6), for one hidden variable. The summation sign over $\{e\}$ denotes marginalization over all evidence variables. Because, for a representative database, the prior $p(y^{(0)})$ equals the frequency of the instantiation $y^{(0)}$ in $\bar{Y}$.

```
(B.3)
```

Otherwise, if $y^{(0)} \neq y^{(0)}$ for some $i$, then $p(y^{(0)} | \bar{y}^{(0)}) > p(y^{(0)} | \bar{y}^{(0)})$, by definition (B.1). It follows that every term in the summation (B.2) is either equal to or greater than the corresponding term (i.e., with the same $\{e\}$) in the first summation in (B.3). Thus, $Q_{\text{BN}}(Y) \geq Q_{\text{ET}}(Y | \bar{Y})$. Since the latter inequality holds for any hidden variable $Y \in \bar{Y}$, the following inequality also holds:

\[
m \cdot Q_{\text{BN}}(\bar{Y}) = \sum_{i=1}^{m} \sum_{j=1}^{N_e} p(\gamma_i | y^{(0)}) p(y^{(0)}) \geq \sum_{i=1}^{m} \sum_{j=1}^{N_e} p(\gamma_i | y^{(0)}) p(y^{(0)}) = m \cdot Q_{\text{ET}}(Y | \bar{Y}).
\]

Thus, $Q_{\text{BN}}(\bar{Y}) \geq Q_{\text{ET}}(Y | \bar{Y})$. (B.4)

References