Knowledge-Based Systems xxx (2008) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Constructing Bayesian networks for criminal profiling from limited data

K. Baumgartner [a,1], S. Ferrari [a,*,2], G. Palermo [b,3]

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

ARTICLE INFO

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

Keywords:
Criminal profiling
Crime analysis
Automation
Bayesian networks
Performance metrics

ABSTRACT

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%.

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 *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

* Corresponding author. Tel.: +1 919 660 5484; fax: +1 919 660 8963.
E-mail address: sferrari@duke.edu (S. Ferrari).
[1] K. Baumgartner is a graduate student in the Pratt School of Engineering.
[2] S. Ferrari is with Faculty of Mechanical Engineering and Material Science and the Faculty of Electrical and Computer Engineering.
[3] G. Palermo is with Faculty of Psychiatry and Neurology.

0950-7051/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2008.03.019

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 and 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 $\Lambda = \{\mathscr{X}, \mathscr{S}\}$ is composed of a set of directed arcs $\mathscr{S}$ that represent the dependencies among a set of variables or nodes $\mathscr{X} = \{X_1, \ldots, X_n\}$ known as *universe*, such that $\mathscr{S} = \{(X_i, X_j) | X_i, X_j \in \mathscr{X}, X_i \neq X_j, j > i\}$. 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 $X_i = x_{i,j}$. $\Theta = \{\theta_1, \ldots, \theta_n\}$ is the parameter structure that is attached to Λ, where θ_i is the conditional probability $p(X_i|\pi_i)$ attached to node X_i, and the set π_i represents the immediate parents of X_i.

In this research, the nodes $\mathscr{X}$ and their instantiations are defined by criminologists and psychologists. The BN arcs $\mathscr{S}$ and parameters Θ are learned from the database $\mathscr{T}$ 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 $\mathscr{S}$ given $\mathscr{T}$, $p(\mathscr{S}|\mathscr{T})$ [18]. Since $p(\mathscr{T})$ is independent of $\mathscr{S}$, the joint probability $p(\mathscr{S}, \mathscr{T})$ can be maximized in place of $p(\mathscr{S}|\mathscr{T})$. A tractable scoring metric, known as *K*2, is obtained from $p(\mathscr{S}, \mathscr{T})$ using the assumptions in [6], which include fixed ordering of variables in $\mathscr{X}$:

$$\mathscr{G} = \log\left(\prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(\overline{N}_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}!\right), \tag{1}$$

where r_i is the number of possible instantiations of X_i, and q_i is the number of unique instantiations of π_i. N_{ijk} is the number of cases in $\mathscr{T}$ in which $X_i = x_{i,k}$, and $\overline{N}_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. 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 (Λ, Θ) represents a factorization of the joint probability over a discrete sample space,

$$p(\mathscr{X}) = p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i|\pi_i), \tag{2}$$

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, $\mu_i \subset \mathscr{X}$, that are conditionally dependent on X_i,

$$p(X_i|\bar{\mu}_i) = \frac{p(\bar{\mu}_i|X_i)p(X_i)}{p(\bar{\mu}_i)}, \tag{3}$$

where $p(X_i)$ is the prior probability of X_i. The likelihood function is factored as $p(\bar{\mu}_i|X_i) = \prod_j p(\bar{\mu}_{i(j)}|X_i)$, where $\bar{\mu}_{i(j)}$ is the evidence of the j^{th} variable in μ_i. The marginalization required to obtain $p(\bar{\mu}_i)$ is simplified using (2):

$$p(\bar{\mu}_i) = \sum_{i=1}^{n} p(\bar{\mu}_i, X_i) = \sum_{k=1}^{r_i} p(X_i = x_{i,k}) \prod_{j=1}^{l} p(\bar{\mu}_{i(j)}|X_i), \tag{4}$$

The posterior probability in (3) is used to obtain the prediction $\overline{X_i} = \arg\max_k p(X_i = x_{i,k}|\bar{\mu}_i)$, 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 $\mathscr{X}$ 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 $X_i \in \mathscr{X}$ is binary, and represents a characteristic or event that is either present or absent at the crime. $\mathscr{X}$ is partitioned into set $\mathscr{E} = \{E_1, \ldots, E_k\}$ containing $k = 36$ *evidence variables* that are observable from the crime scene, and set $\mathscr{Y} = \{Y_1, \ldots, Y_m\}$ containing $m = 21$ *offender variables* that characterize the offender and, thus, are unknown or *hidden* at the crime scene.

The BN model structure, $\mathscr{S}$, and parameters, Θ, are learned using a police database of cleared single-victim single-offender homicides, $\mathscr{D} = \{C_1, \ldots, C_d\}$. 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

Table 1
Definition of selected offender and crime scene variables

Variable	Definition
Y_4	Prior record of property damage
Y_5	Prior record of disorderly conduct
Y_6	Previous imprisonment or youth detention
Y_9	History of sex crime
Y_{10}	Record of armed services
E_{11}	Victim sustained stabbing wounds
E_{12}	Blunt instrument used on victim
E_{13}	Offender used own body as weapon (e.g. strangulation)
E_{14}	Victim was shot
E_{15}	Victim sustained wounds to head (excluding face and neck)
E_{16}	Victim sustained wounds to face (ears forward)
E_{31}	Victim was sexually assaulted
E_{33}	Arson to crime scene or body
E_{34}	Body was found in water

of $\mathscr{X}$ that is obtained from a cleared homicide case. The values of all variables in $\mathscr{Y}$ are obtained from interviews with the convicted offender, and the values of all variables in $\mathscr{E}$ are obtained from the police record of the investigation. Then, $\mathscr{D}$ is randomly partitioned into a training set $\mathscr{T}$ and a validation set $\mathscr{V}$, such that $\mathscr{T} \cap \mathscr{V} = \emptyset$. 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 $\mathscr{T}$ 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 $\mathscr{V}$ in order to determine the accuracy of the BN predictions. However, since $\mathscr{T} \cap \mathscr{V} = \emptyset$, the same approach and accuracy would apply to new unsolved cases, such as those encountered by investigators.

Let $\bar{e} = \{\bar{e}_1, \ldots, \bar{e}_k\}$ denote the evidence obtained from the crime scene of a new case, where $\bar{e}_i$ 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 *jtree_inf_engine* [29], is used to compute the posterior probability distribution for every offender variable $Y_i \in \mathscr{Y}$. An offender variable prediction, $\hat{Y}_i$, is the instantiation with the largest posterior probability, $\hat{Y}_i = y_i^* \equiv \arg\max_\ell p(Y_i = y_{i,\ell}|\bar{e})$, and $p(y_i^*|\bar{e})$ is said to be the confidence level (CL) of the prediction $\hat{Y}_i$. The set of all predictions, $\hat{\mathscr{Y}} = \{\hat{Y}_1, \ldots, \hat{Y}_m\}$, 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 $\mathscr{Y}$ are predicted correctly from the evidence $\bar{e}$, 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 $\mathscr{X}$, then its predictive accuracy for m hidden variables is,

$$Q_{\mathrm{BN}}(\mathscr{Y}) = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{N_e}\max_\ell[p(y_{i,\ell})p(\xi_j|y_{i,\ell})], \tag{5}$$

where ξ_j denotes the simultaneous occurrence of a set of instantiations of $\mathscr{E}$, and thus $j = 1, \ldots, N_e$, with $N_e = \prod_k r_i$. The proof and a simple example are provided in Appendix A. As the size of $\mathscr{T}$ becomes sufficient and the trained BN approaches the actual joint probability distribution, this learning curve approaches Q_{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 $\mathscr{D}$, and denote it by $\hat{Y}_i^F$. If Y_i is independent of any other variable in $\mathscr{X}$, then $\hat{Y}_i^F$ is the optimal prediction, and the average accuracy for m hidden variables is

$$Q_{FO}(\mathscr{Y}|\mathscr{D}) = \frac{1}{m}\sum_{i=1}^{m}\max_\ell f(\bar{y}_{i,\ell}|\mathscr{D}), \tag{6}$$

where the frequency $f(\bar{y}_{i,\ell}|\mathscr{D})$ is the number of cases in $\mathscr{D}$ in which $Y_i = y_{i,\ell}$ divided by d, and $(\bar{\cdot})$ denotes evidence of the instantiation. It is shown in Appendix B that if the variables in $\mathscr{Y}$ are not independent and the BN accurately represents the joint probability distribution over $\mathscr{X}$, then $Q_{\mathrm{BN}} \geqslant Q_{FO}$. It follows that when a trained BN displays $Q_{\mathrm{BN}} < Q_{FO}$, it cannot represent the underlying probability distribution and, thus, the database $\mathscr{T}$ is insufficient. Therefore, $Q_{\mathrm{BN}} \geqslant Q_{FO}$ is a necessary but not sufficient condition for deeming a database $\mathscr{T}$ sufficient for training.

When $Q_{\mathrm{BN}} \geqslant Q_{FO}$, the structural robustness of the trained BN model can provide additional insight into the sufficiency of $\mathscr{T}$. A BN defined over a fixed universe $\mathscr{X}$ is said to be *structurally robust* if its arc structure is insensitive to small changes in the training set $\mathscr{T}$. 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,

$$S = \begin{bmatrix} s_{1,1} & \cdots & s_{1,c} & \cdots & s_{1,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ s_{n,1} & \cdots & s_{n,c} & \cdots & s_{n,n}, \end{bmatrix} \tag{7}$$

where each entry $s_{r,c}$ represents the presence (±1) or absence (0) of an arc between two variables $X_r, X_c \in \mathscr{X}$. The direction of the arcs is positive according to the expert node ordering and, thus, $s_{r,c} = +1$ (and, $s_{c,r} = -1$), when $r < c$ and there exist an arc $X_r \to X_c$. Let $\mathscr{T}$ represent an available training set, and $\mathscr{T}'$ represent $\mathscr{T}$ 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 $\mathscr{S}$ and $\mathscr{S}'$, respectively, such that:

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

$$\Delta A_{10} = \sum_{r=1}^{n}\sum_{c=1}^{n} |s_{r,c} - s'_{r,c}|. \quad (8)$$

Thus, the metric $1/\Delta A_{10}$ can be used to represent BN structural robustness, and is expected to increase as the size of $\mathscr{T}$ increases. A lack of structural robustness typically indicates that $\mathscr{S}$ is not reliable, and that the parameters Θ learned subsequently to $\mathscr{S}$ may also be inaccurate. Therefore, when the size of $\mathscr{T}$ 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 $\mathscr{Y}$ brought about by evidence of $\mathscr{E}$. The mutual information between two sets of random variables, $I(\mathscr{Y};\mathscr{E})$, represents the amount of information that $\mathscr{E}$ has about $\mathscr{Y}$, and is related to the entropy as follows [7]:

$$I(\mathscr{Y};\mathscr{E}) = H(\mathscr{Y}) - H(\mathscr{Y}|\mathscr{E}) = H(\mathscr{Y}) + H(\mathscr{E}) - H(\mathscr{Y},\mathscr{E}). \quad (9)$$

Thus, we let $I(\mathscr{Y};\mathscr{E})$ denote the *mutual information of a Bayesian network* with hidden variables $\mathscr{Y}$ to be inferred from evidence of the variables $\mathscr{E}$. The *joint entropy* of the BN can be computed using the BN factorization (2):

$$H(\mathscr{Y},\mathscr{E}) = H(\mathscr{X}) = H(X_1,\ldots,X_n) = \sum_{i=1}^{n} H(X_i|\pi_i). \quad (10)$$

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,

$$\begin{aligned} I(\mathscr{Y};\mathscr{E}) &= \sum_{j=1}^{k}[H(E_j) - H(E_j|\pi_j)] + \sum_{i=1}^{m}[H(Y_i) - H(Y_i|\pi_i)] \\ &= \sum_{j=1}^{k}\sum_{\ell=1}^{r_j}\Bigg\{\sum_{\pi_j} p(\pi_j)p(e_{j,\ell}|\pi_j)\log[p(\pi_j)p(e_{j,\ell}|\pi_j)] \\ &\quad - p(\pi_j)p(e_{j,\ell}|\pi_j)\log[p(\pi_j)p(e_{j,\ell}|\pi_j)]\Bigg\} \\ &\quad + \sum_{i=1}^{m}\sum_{\ell=1}^{r_i}\Bigg\{\sum_{\pi_i} p(\pi_i)p(y_{i,\ell}|\pi_i)\log[p(\pi_i)p(y_{i,\ell}|\pi_i)] \\ &\quad - p(\pi_i)p(y_{i,\ell}|\pi_i)\log[p(\pi_i)p(y_{i,\ell}|\pi_i)]\Bigg\}, \end{aligned} \quad (11)$$

where the summations over $\pi_i, \pi_j \in \mathscr{X}$ denote marginalization. As can be seen from (11), $I(\mathscr{Y};\mathscr{E})$ 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 $\mathscr{T}$ is available, several databases of increasing size can be obtained by sampling $\mathscr{T}$. 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_{FO} to determine whether $\mathscr{T}$ 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 $\mathscr{T}$ with 5000 cases is sufficient for obtaining the BN model of criminal behavior described in Section 3. The universe $\mathscr{X}$ 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_{BN}, $1/\Delta A_{10}$, and $I(\mathscr{Y};\mathscr{E})$, 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.

It is well known that structural training can be significantly improved by impeding arcs *a priori* based on expert knowledge [13] and heuristics [14,16]. We present a new and simple approach for deciding which arcs to inhibit that is based on Markov separation properties. This approach, referred to as K2′, is applicable to BNs in which evidence is always available about the same subset of variables, $\mathscr{E}$. Markov separation properties are typically exploited to simplify inference (Section 2). In the K2′ algorithm these properties are exploited to simplify structural training by inhibiting arcs between the evidence variables in $\mathscr{E}$. When all variables in $\mathscr{E}$ are instantiated, a BN structure can be Markov equivalent to one in which these arcs are removed, and thus can produce the same inference results even if the evidence variables are not independent. In particular, any two evidence variables $E_i, E_j \in \mathscr{E}$ are d-separated $E_i \perp E_j$, [22], if for all paths between them there is an intermediate variable, X_l, such that the connection is (*I*) diverging and $X_l \in \mathscr{E}$, or (*II*) the connection is converging, and $X_l, \mu_l \in \mathscr{Y}$. If these conditions do not apply, a small error is introduced only if $E_i \not\perp E_j$. Both instances are illustrated by an example in Fig. 1. The numerical results presented in this section demonstrate that BNs obtained by the K2′ outperform the BNs obtained by the *K*2 algorithm using the same training data.

The predictive accuracy of the criminal profiling BN is plotted on the learning curve in Fig. 2. It can be seen that for a BN trained

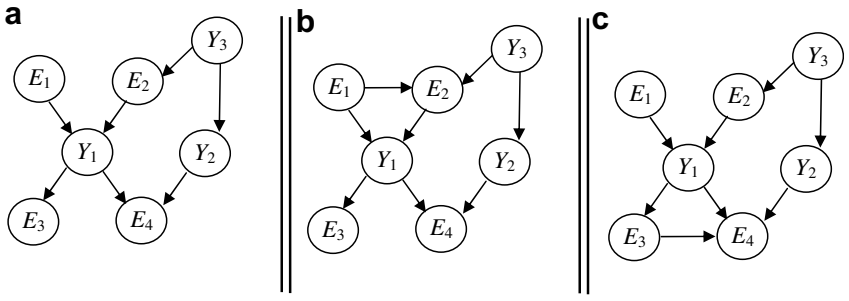


Fig. 1. A simple BN structure learned by the K2′ algorithm (a) omitting arcs between evidence variables is compared to a K2 structure that is Markov equivalent (b), and to one that is not (c).

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

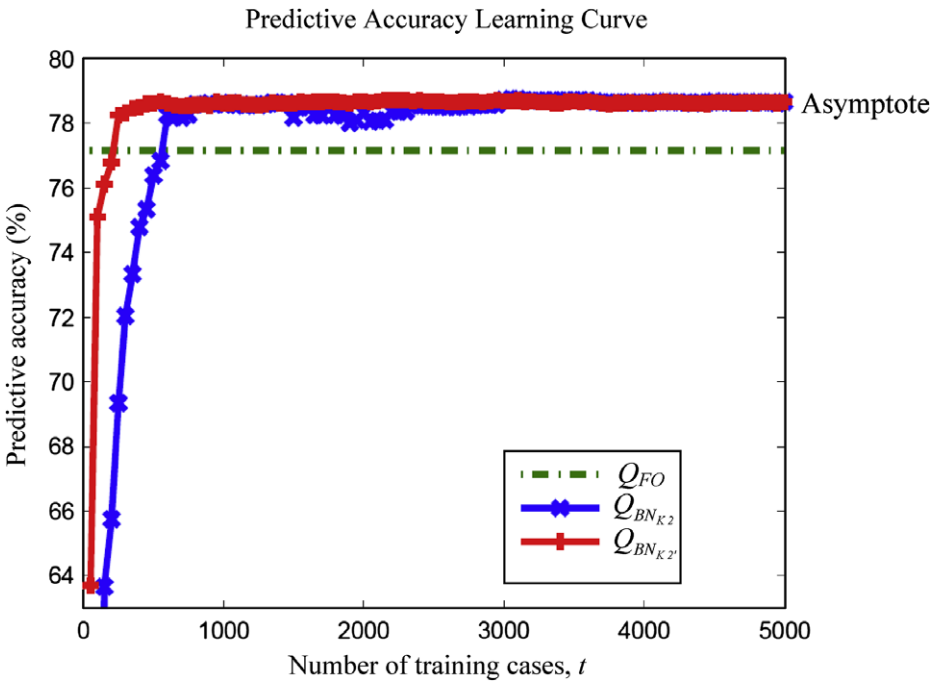


Fig. 2. Predictive accuracy learning curve for BNs obtained by the K2′ and K2 algorithms, with $\Delta t = 50$.

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 $\mathscr{T}$. Then, the structural robustness metric ΔA_{10} in (8) is plotted on the learning curve in Fig. 3, where the abscissa represents the size t before $0.01 \cdot t$ cases are removed. As in Fig. 2, the training sets sampled from $\mathscr{T}$ have a size t that varies between 50 and 5000 in increments of 50. When the training set contains more than 1000 cases, $0 \leqslant \Delta A_{10} < 10$ 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

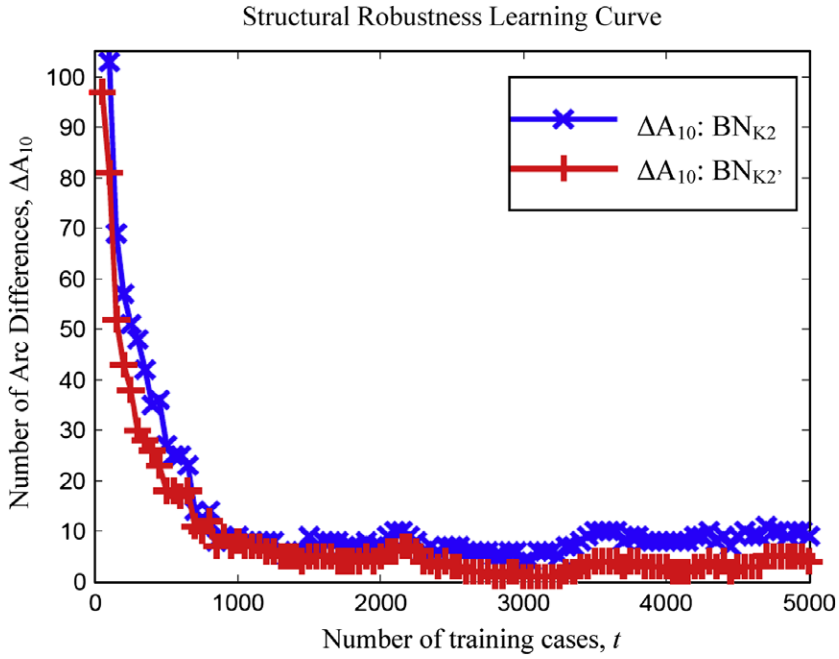


Fig. 3. Structural robustness learning curve for BNs obtained by the K2′ and K2 algorithms, with $\Delta t = 50$.

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

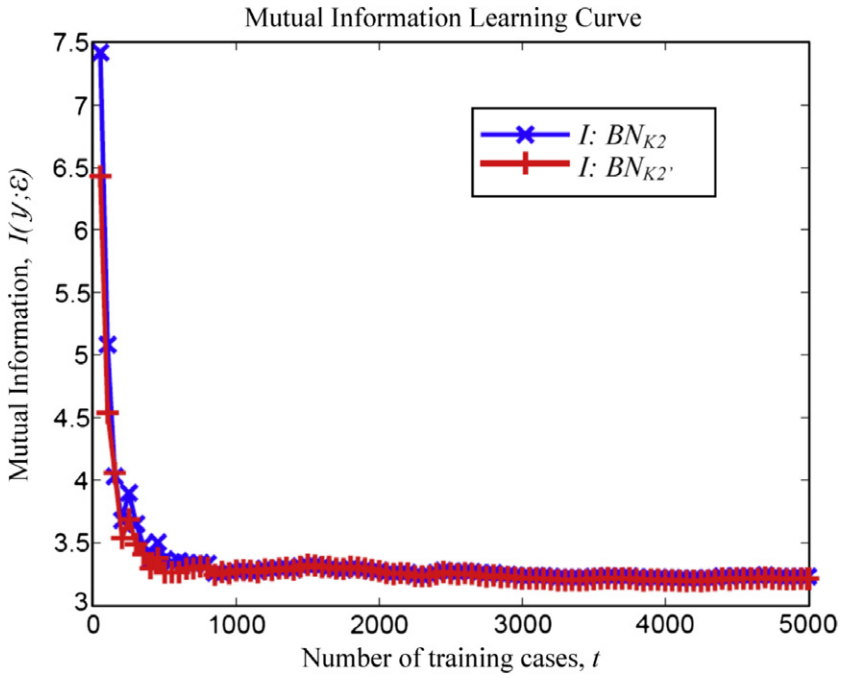


Fig. 4. Mutual information learning curve for BNs obtained by the K2′ and K2 algorithms, with $\Delta t = 50$.

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 $\mathcal{V}$ 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 $\mathcal{Y}$. 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-

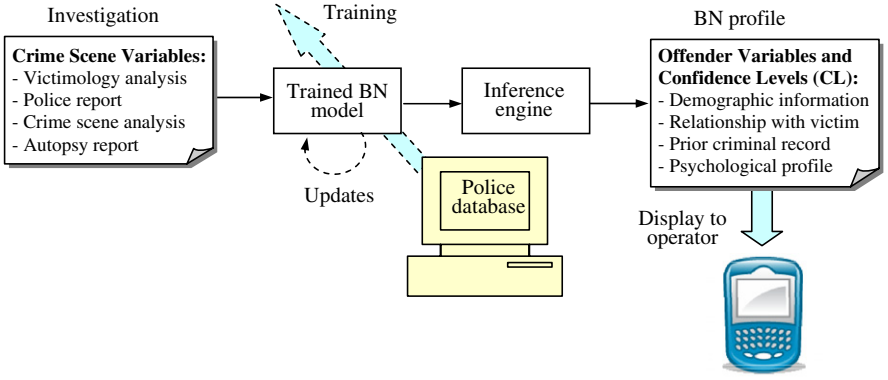


Fig. 5. Implementation of BN model as decision-support tool for police investigations.

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

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

Offender profile	First homicide		
Variables	Actual	BN (CL %)	Expert team
Y_3 = prior record of violence	No	No (74)	No
Y_4 = prior record of damage	No	No (84)	No
Y_8 = unemployed	Yes	No (51)	Yes
Y_{11} = familiar with CS	Yes	Yes (84)	Yes
Y_{12} = gender	Male	Male (73)	Male
Y_{14} = prior record of abuse	No	No (80)	Yes
Y_{17} = prior record of fraud	No	No (67)	No
Y_{19} = prior sexual relationship with victim	No	No (57)	Yes
Y_{20} = blood relative of victim	No	No (90)	Yes\No
Average accuracy (CL %)	–	90.5% (79)	62%

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, $\widehat{Y}_8$, compared to three incorrect predictions by the experts. Even more importantly, $\widehat{Y}_8$ 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 $\mathscr{D}$). 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 $\mathscr{V}$ 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

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

Offender profile	Second homicide		
Variables	Actual	BN (CL %)	Expert team
Y_3 = prior record of violence	No	No (83)	Yes\No
Y_4 = prior record of damage	No	No (95)	Yes\No
Y_8 = unemployed	No	Yes (66)	Yes\No
Y_{11} = familiar with CS	Yes	Yes (88)	Yes
Y_{12} = gender	Female	Female (99)	Female
Y_{14} = prior record of abuse	No	No (80)	Yes
Y_{17} = prior record of fraud	No	No (80)	Yes
Y_{19} = prior sexual relationship with victim	No	No (99)	Yes
Y_{20} = blood relative of victim	Yes	No (66)	Yes\No
Average accuracy (CL %)	–	85.7% (86)	47.6%

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

Confidence level (%)	Correct ‖ Total (number of predictions)	Average accuracy (%)
CL $\geqslant$ 50	16,511 ‖ 21,000	78.6
CL $\geqslant$ 60	15,054 ‖ 18,361	82
CL $\geqslant$ 70	12,845 ‖ 14,952	85.9
CL $\geqslant$ 80	10,515 ‖ 11,802	89.1
CL $\geqslant$ 90	5,084 ‖ 5,321	95.6

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 $\geqslant$ 70%, then its accuracy is approximately 85.9% (Table 4). Also, if an offender variable, such as Y_{11}, displays a confidence level 80% $\leqslant$ 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 $\mathscr{X}$, 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-

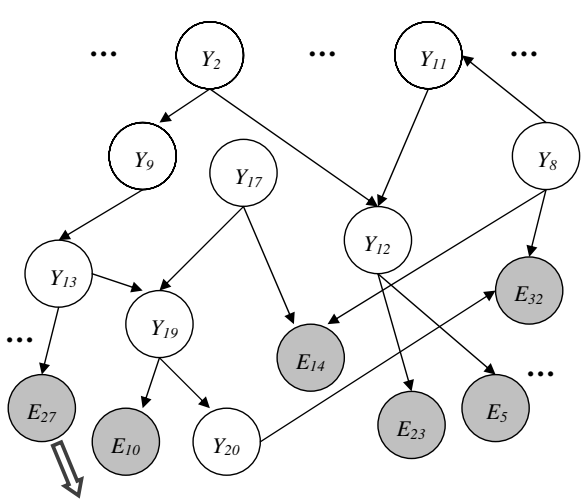
Table 5
Accuracy range for offender variables that display, on average, a confidence level within the specified ranges

Confidence level range (%)	Accuracy range (%)	Number of offender variables (Example)
50–60	55.8	1, (Y_8 = unemployed)
60–70	63.8–67.8	4, (Y_{19} = priori sexual relationship)
70–80	71.0–77.7	5, (Y_5 = prior record of burglary)
80–90	79.4–92.4	8, (Y_{11} = familiarity with crime scene)
90–100	97.0–98.8	3, (Y_{20} = blood relative of victim)

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

Offender variable	Average percent accuracy
Y_{19} = sexual relationship with victim	63.8%
Y_{16} = prior record of fraud	67.7%
Y_6 = history of psychiatric/social problems	67.8%
Y_{17} = prior record of fraud	71.0%
Y_3 = prior record of violence	73.6%
Y_5 = prior record of burglary	77.6%
Y_4 = prior record of damage	77.7%
Y_{14} = prior record of abuse	79.4%
Y_7 = young offender, 17–21 years old	82.8%
Y_{11} = familiarity with crime scene	86.1%
Y_{12} = male gender	89.7%
Y_{20} = blood relative to victim	91.1%

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



Offender Variables
Y_2 = Prior record of burglary
Y_8 = Unemployed
Y_9 = Prior record of sex crime
Y_{11} = Familiar with crime scene
Y_{12} = Gender
Y_{13} = Knew victim
Y_{17} = Prior record of fraud
Y_{19} = Prior sexual relationship w/ victim
Y_{20} = Blood relative of victim

Crime Scene Variables
E_5 = Victim's face not deliberately hidden
E_{10} = Blindfolded
E_{23} = Weapon from scene
E_{14} = Victim was shot
E_{27} = Body hidden outdoors
E_{32} = Suffocation

$p(E_{27}\|Y_{13})$	$p(e_{27,1}\|Y_{13})$	$p(e_{27,2}\|Y_{13})$
$Y_{13} = y_{13,1}$	0.05	0.95
$Y_{13} = y_{13,2}$	0.32	0.68

Fig. 6. Slice of the trained BN structure, including one example of a CPT attached to node E_{27}.

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_{19}) 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_i, and let $\mathscr{U} = \{Y_j | Y_j \in \mathscr{Y}, j \neq i\}$. The prediction $\widehat{Y}_i$ is obtained by inferring Y_i from evidence $\bar{e}$ such that,

$$\widehat{Y}_i = y_i^* \equiv \arg\max_{\ell} p(y_{i,\ell}|\bar{e}) = \arg\max_{\ell} p(y_{i,\ell}, \bar{e}), \quad \ell = 1, \ldots, r_i. \tag{A.1}$$

Therefore, the predictive accuracy of one hidden variable is,

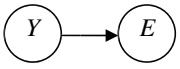
$$Q_{\mathrm{BN}}(Y_i) = p(y_i^*|\bar{e})p(\bar{e}), \tag{A.2}$$

and $Q_{\mathrm{BN}} \in [0,1]$ for any r_i. Since there are $N_e = \prod_k r_i$ possible evidence combinations, the probability of any simultaneous occurrence of a set of instantiations $\xi_j \equiv \{e_{1,j} \cup \ldots \cup e_{k,j}\}$ is the prior $p(\xi_j)$, 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,

$$Q_{\mathrm{BN}}(Y_i) = \sum_{j=1}^{N_e} \max_{\ell} p(y_{i,\ell}|\xi_j)p(\xi_j) = \sum_{j=1}^{N_e} \max_{\ell} p(\xi_j|y_{i,\ell})p(y_{i,\ell}), \tag{A.3}$$

where $p(y_{i,\ell}) = \sum_{\pi_i} p(\pi_i)p(y_{i,\ell}|\pi_i)$. Then, for a set of m hidden variables, $\mathscr{Y} = \{Y_1, \ldots, Y_m\}$, the BN predictive accuracy is the average of the individual predictive accuracies

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



$p(Y)$:

$p(Y=y_1)$	$p(Y=y_2)$
0.7	0.3

$p(E|Y)$:

	$p(e_1\|Y)$	$p(e_2\|Y)$
y_1	0.571	0.429
y_2	0.667	0.333

Fig. 7. Example of a two-node Bayesian network model.

$$Q_{BN}(\mathscr{Y}) = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{N_e}\max_{\ell}[p(y_{i,\ell})p(\xi_j|y_{i,\ell})] = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{N_e}\max_{\ell}\left[p(\xi_j|y_{i,\ell})\sum_{\pi_i}p(y_{i,\ell}|\pi_i)p(\pi_i)\right]. \quad (A.4)$$

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,

$$Q_{BN}(Y) = \sum_{j=1}^{N_e=2}\max_{\ell}p(y_\ell|\xi_j)p(\xi_j) = \sum_{j=1}^{N_e=2}\max_{\ell}p(\xi_j|y_\ell)p(y_\ell), \quad (A.5)$$

where $\ell = 1,2$, $\xi_1 = \{e_1\}$, and $\xi_2 = \{e_2\}$. Hence, for the BN in Fig. 7, $Q_{BN}(Y) = 0.7$.

Appendix B. Frequency of occurrence lower bound

Consider one hidden variable Y_i and a set of k evidence variables $\mathscr{E}$. Then, the predictive accuracy of a prediction $\widehat{Y}_i$ is given by (A.3), where, $\ell = 1,\ldots,r_i$, and N_e is the number of all possible evidence combinations. Similarly to (A.1), based on the definition of prediction we let $y_i^{*(j)}$ denote the instantiation with the highest posterior probability when $\mathscr{E} = \xi_j$, namely,

$$y_i^{*(j)} \equiv \arg\max_{\ell}\{p(y_{i,l}|\xi_j)\} = \arg\max_{\ell}\{p(y_{i,l}|\xi_j)p(\xi_j)\} = \arg\max_{\ell}\{p(\xi_j|y_{i,l})p(y_{i,l})\} \quad (B.1)$$

using Bayes' rule. It follows that $p(\xi_j|y_i^{*(j)})p(y_i^{*(j)}) \geqslant p(\xi_j|y_{i,l})p(y_{i,l})$ $\forall \ell = 1,\ldots,r_i$. Also (A.3) can be written as

$$Q_{BN}(Y_i) = \sum_{j=1}^{N_e}p(\xi_j|y_i^{*(j)})p(y_i^{*(j)}). \quad (B.2)$$

Now let $y_i^{*(F)}$ denote the instantiation chosen as the frequency of occurrence prediction $\widehat{Y}_i^F$, such that $y_i^{*(F)} = \arg\max_\ell f(\bar{y}_{i,\ell}|\mathscr{D})$, where both instantiations $y_i^{*(j)}$ and $y_i^{*(F)}$ belong to the domain of Y_i, which contains r_i mutually exclusive instantiations denoted by $y_{i,\ell}$ (Section 2). Then, if $y_i^{*(j)} = y_i^{*(F)}$ $\forall j = 1,\ldots,N_e$, and the database $\mathscr{D}$ is representative of the statistical population, the predictive accuracy (B.2) simplifies to,

$$Q_{BN} = \sum_{j=1}^{N_e}p(\xi_j|y_i^{*(F)})p(y_i^{*(F)}) = \sum_{\mathscr{E}}p(\mathscr{E}|y_i^{*(F)})p(y_i^{*(F)}) = p(y_i^{*(F)}) = f(y_i^{*(F)}|\mathscr{D}) = \max_{\ell}f(\bar{y}_{i,\ell}|\mathscr{D}) = Q_{FO}(Y_i|\mathscr{D}) \quad (B.3)$$

achieving the lower bound Q_{FO} in (6), for one hidden variable. The summation sign over $\mathscr{E}$ denotes marginalization over all evidence variables. Because, for a representative database, the prior $p(y_i^{*(F)})$ equals the frequency f of the instantiation $y_i^{*(F)}$ in $\mathscr{D}$.

Otherwise, if $y_i^{*(j)} \neq y_i^{*(F)}$ for some j, then $p(\xi_j|y_i^{*(j)})p(y_i^{*(j)}) > p(\xi_j|y_i^{*(F)})p(y_i^{*(F)})$, 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 j) in the first summation in (B.3). Thus, $Q_{BN}(Y_i) \geqslant Q_{FO}(Y_i|\mathscr{D})$. Since the latter inequality holds for any hidden variable $Y_i \in \mathscr{Y}$, the following inequality also holds:

$$m \cdot Q_{BN}(\mathscr{Y}) = \sum_{i=1}^{m}\sum_{j=1}^{N_e}p(\xi_j|y_i^{*(j)})p(y_i^{*(j)}) \geqslant \sum_{i=1}^{m}\sum_{j=1}^{N_e}p(\xi_j|y_i^{*(F)})p(y_i^{*(F)}) = m\cdot\sum_{i=1}^{m}Q_{FO}(Y_i|\mathscr{D}) = m\cdot Q_{FO}(\mathscr{Y}|\mathscr{D}) \quad (B.4)$$

Thus, $Q_{BN}(\mathscr{Y}) \geqslant Q_{FO}(\mathscr{Y}|\mathscr{D})$.

References

[1] B. Abramson, ARCO1: an application to the oil market, Proceedings of the Conference of Artificial Intelligence (1991) 1–8.
[2] L. Alison, C. Bennell, D. Ormerod, The personality paradox in offender profiling: a theoretical review of the process involved in deriving background characteristics from crime scene actions, Psychology, Public Policy, and Law 8 (1) (2002) 115–135.
[3] K.A.C. Baumgartner, S. Ferrari, C.G. Salfati, Bayesian network modeling of offender behavior for criminal profiling, Proceedings of the Conference of Decision and Control (2005) 2702–2709.
[4] J.W. Brahan, K.P. Lam, H. Chan, W. Leung, AICAMS: artificial intelligence crime analysis and management system, Knowledge-Based Systems (1998) 355–361.
[5] W. Buntine, A guide to the literature on learning probabilistic networks from data, IEEE Transactions on Knowledge Data Engineering 8 (2) (1996) 195–210.
[6] G.F. Cooper, E. Herskovits, A Bayesian method for the induction of probabilistic networks from data, Machine Learning 9 (1992) 309–347.
[7] T.M. Cover, J.A. Thomas, Elements of Information Theory, John Wiley and Sons Inc., 1991.
[8] R. Cowell, Advanced inference in Bayesian networks, in: M. Jordan (Ed.), Learning in Graphical Models, 1998, pp. 27–50.
[9] R. Cowell, Introduction to inference for Bayesian networks, in: M. Jordan (Ed.), Learning in Graphical Models, 1998, pp. 9–26.
[10] R. Cowell, A. Dawid, S. Lauritzen, D. Spiegelhalter, Probabilistic Networks and Expert Systems, Springer, 1999.
[11] D. Dash, G. Cooper, Model averaging for predictions with discrete Bayesian networks, Journal of Machine Learning Research 5 (2004) 1177–1203.
[12] S.A. Egger, Psychological profiling: past, present, and future, Journal of Contemporary Criminal Justice 15 (3) (1999) 242–261.
[13] S. Ferrari, A. Vaghi, Demining sensor modeling and feature-level fusion by Bayesian networks, IEEE Sensors 6 (2) (2006) 471–483.
[14] N. Friedman, D. Geiger, Bayesian network classifiers, Machine Learning 29 (1997) 131–163.
[15] N. Friedman, D. Koller, Being Bayesian about network structure: a Bayesian approach to structure discovery in Bayesian networks, Machine Learning 50 (2003) 95–125.
[16] N. Friedman, I. Nachman, D. Peer, Learning Bayesian network structure from massive datasets: the "sparse candidate" algorithm, UAI 29 (1999) 206–215.
[17] P. Gottschalk, Stages of knowledge management systems in police investigations, Knowledge-Based Systems 19 (2006) 381–387.
[18] D. Heckerman, A Bayesian approach to learning causal networks, Technical Report MSR-TR-95-04, May 1995, pp. 1–23.
[19] D. Heckerman, J. Breese, B. Nathwani, Toward normative expert systems I: the PATHFINDER project, Methods of Information Medicine 31 (1992) 90–105.
[20] M. Hohenner, S. Wachsmuth, G. Sagerer, Modelling expertise for structure elucidation in organic chemistry using Bayesian networks, Knowledge-Based Systems 18 (4–5) (2005) 207–215.
[21] K.-B. Hwang, B.-T. Zhang, Bayesian model averaging of Bayesian networks classifiers over multiple node-orders: Application to sparse datasets, IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics 35 (6) (2005) 1302–1310.
[22] F.V. Jensen, Bayesian Networks and Decision Graphs, Springer-Verlag, 2001.
[23] A. Joseph, N. Fenton, M. Neil, Predicting football results using Bayesian nets and other machine learning techniques, Knowledge-Based Systems 19 (7) (2006) 544–553.
[24] R.N. Kocsis, Criminal Profiling: Principles and Practice, Humana Press, 2006.
[25] R.N. Kocsis, R.W. Cooksey, H.J. Irwin, Psychological profiling of offender characteristics from crime scene behaviors in serial rape offences, International Journal of Offender Therapy and Comparative Criminology 46 (2) (2002) 144–169.
[26] P. Larranaga, C.H. Kuijpers, R.H. Murga, Y. Yurramendi, Learning Bayesian network structures by searching for the best ordering with genetic algorithms, IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans 26 (4) (1996) 487–493.
[27] M. Lefly, V. Austin, Match'em: Using fuzzy logic to profile criminals, Proceedings of the Sixth IEEE International Conference on Fuzzy Systems 1 (1) (1997) 305–311.

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

[28] J. Liu, K. Chang, J. Zhou, Learning Bayesian networks with a hybrid convergent method, IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans 29 (5) (1999) 436–449.
[29] K. Murphy, How To Use Bayes Net Toolbox, 2004 (Online). Available from: <http://www.ai.mit.edu/murphyk/Software/BNT/bnt.html>.
[30] D. Nikovski, Constructing Bayesian networks for medical diagnosis from incomplete and partially correct statistics, IEEE Transaction on Knowledge and Data Engineering 12 (4) (2000) 509–516.
[31] G. Palermo, R. N. Kocsis, in: Charles C. Thomas (Ed.), Offender Profiling: An Introduction to the Sociopsychological Analysis of Violent Crime, Springfield, IL, 2004.
[32] A.J. Pinizzotto, Finkel, criminal personality profiling: an outcome and process study, Law and Human Behavior 14 (3) (1990) 215–233.
[33] M. Reidmiller, H. Braun, A direct adaptive method for faster backpropagation learning: the RPROP algorithm, Proceedings of the IEEE International Conference on NN (ICNN) (1993) 586–591.
[34] R.K. Ressler, A. Burgess, J.E. Douglas, Sexual Homicide: Patterns and Motives, Lexington Books, New York, 1988.
[35] D. Rossmo, Geographical Profiling: Target Patterns of Serial Murderers, Doctoral dissertation, Simon Frasier University, Vancouver, British Columbia, Canada, 1995.
[36] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, second ed., Prentice Hall, 2002.
[37] C. Salfati, Profiling homicide: a multidimensional approach, Homicide Studies 4 (2000) 265–293.
[38] C.G. Salfati, Offender interaction with victims in homicide: a multidimensional analysis of crime scene behaviors, Journal of Interpersonal Violence 18 (5) (2003) 490–512.
[39] C.G. Salfati, D.V. Canter, Differentiating stranger murders: profiling offender characteristics from behavioral styles, Behavioral Science and the Law 17 (1999) 391–406.
[40] G. Santafé, J. Lozano, P. Larrañaga, Bayesian model averaging of naive Bayes for clustering, IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics 36 (5) (2006) 1149–1161.
[41] P. Santtila, H. Häkkänen, D. Canter, T. Elfgren, Classifying homicide offenders and predicting their characteristics from crime scene behavior, Scandinavian Journal of Psychology 44 (2003) 107–118.
[42] M. Strano, A neural network applied to criminal psychological profiling: an Italian initiative, International Journal of Offender Therapy and Comparative Criminology 48 (4) (2004) 495–503.
[43] B.E. Turvey, Criminal Profiling: An Introduction to Behavioral Evidence Analysis, second ed., Academic Press, 2002.
[44] M. van Gerven, P. Lucas, Employing maximum mutual information for Bayesian classification, Lecture notes in Computer Science 3337 (2004) 188–199.
[45] S. Yang, K. Chang, Comparison of score metrics for Bayesian network learning, IEEE Transactions on Systems, Man, and Cybernetics 32 (3) (2002) 419–428.

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