J. Intelligent Learning Systems & Applications, **2010, 2, 156-166**
doi:10.4236/jilsa.2010.23019 Published Online August 2010 (http://www.SciRP.org/journal/jilsa)

Scientific Research

Knowledge Discovery for Query Formulation for Validation of a Bayesian Belief Network

Gursel Serpen, Michael Riesen

Electrical Engineering and Computer Science, College of Engineering, University of Toledo; School of Law, University of Toledo, Toledo, USA.
Email: gserpen@eng.utoledo.edu, riesen@fraser-ip.com

Received February 23th, 2010; revised July 6th, 2010; accepted July 20th, 2010.

ABSTRACT

This paper proposes machine learning techniques to discover knowledge in a dataset in the form of if-then rules for the purpose of formulating queries for validation of a Bayesian belief network model of the same data. Although domain expertise is often available, the query formulation task is tedious and laborious, and hence automation of query formulation is desirable. In an effort to automate the query formulation process, a machine learning algorithm is leveraged to discover knowledge in the form of if-then rules in the data from which the Bayesian belief network model under validation was also induced. The set of if-then rules are processed and filtered through domain expertise to identify a subset that consists of "interesting" and "significant" rules. The subset of interesting and significant rules is formulated into corresponding queries to be posed, for validation purposes, to the Bayesian belief network induced from the same dataset. The promise of the proposed methodology was assessed through an empirical study performed on a real-life dataset, the National Crime Victimization Survey, which has over 250 *attributes and well over* 200,000 *data points. The study demonstrated that the proposed approach is feasible and provides automation, in part, of the query formulation process for validation of a complex probabilistic model, which culminates in substantial savings for the need for human expert involvement and investment.*

Keywords: *Rule Induction, Semi-Automated Query Generation, Bayesian Net Validation, Knowledge Acquisition Bottleneck, Crime Data, National Crime Victimization Survey*

1. Introduction

Query formulation is an essential step in the validation of complex probabilistic reasoning models that are induced from data using machine learning or statistical techniques. Bayesian belief networks (BBN) have proven to be computationally viable empirical probabilistic models of data [1]. Advances in machine learning, data mining, and knowledge discovery and extraction fields greatly aided in maturation of Bayesian belief networks, particularly for classification and probabilistic reasoning tasks. A Bayesian belief network can be created through a multitude of means: it can be induced solely from data, handcrafted by a domain expert, or a combination of these two techniques can be leveraged. A Bayesian belief network model essentially approximates the full joint probability distribution in the domain of interest. The development of a Bayesian belief network model is followed by a rigorous validation phase to ascertain that the model in fact approximates the full joint probability distribution reasonably well, even under the set of independence assumptions made. Validation is a comprehensive, multi-part process and often requires costly domain expert involvement and labor.

When a BBN model is used as a probabilistic reasoning engine, the validation requires a complex and challenging approach, wherein a multitude of validation-related activities must be performed [2-5] and as part of one such activity, queries must be formed and posed to the network. Any subset of variables might be considered as evidence in such a query, which leads to the need to formulate an inordinate number of queries based on various subsets of variables. During validation by querying, a value assignment to some variables in the network is made and the posterior marginal probability or expectation of some other variables is desired. In other words, marginal probabilities and expectations can be calculated conditionally on any number of observations or evidence supplied to the network. It is also desirable, given that certain evidence is supplied, to ask for the values of non-evidence variables that result in the maximum possible posterior probability for the evidence, *i.e.*, an *ex-*

planation for the available evidence. One can specify a group of variables in the network to be estimated or estimate all variables in the network collectively. The existing literature for validation of BBNs as probabilistic reasoning tools is sparse and mainly promotes ad hoc approaches or mechanisms.

The formulation of an appropriate "query" requires the use of extrinsic methods in order to discover relationships among attributes. More specifically, in forming a query, access to a specific domain expertise can prove to be an efficient method in choosing which attributes to include as evidence and which attributes to identify for explanation or estimation. Experts in the domain of the focus data can prove to be a useful resource in forming the queries. However, there are many challenges in utilizing domain experts in manual formulation of queries and these challenges are in addition to the shear cost and resources needed.

Conducting interviews with one or preferably more experts in the relevant field of interest is one of the preliminary steps in manual query formulation. Such interviews typically expose many issues and challenges associated with relying on experts in the field to focus and to form queries. Experts interviewed are likely to demonstrate an interest in forming unique queries that would parallel their own expertise or interest, which might not fully overlap with the specific domain on which the model was built [6]. The list of potential queries suggested by the domain experts could prove to be inapplicable as the specific dataset employed to develop the BBN model might not include all the attributes sought by the domain experts. In other circumstances, experts may be interested in applying local and regional attributes rather than the global attributes or the national attributes used in the dataset.

It is highly desirable to develop an automated procedure that formulates queries by leveraging the same dataset that was employed to induce the Bayesian belief network model. In similar terms, exploration of other, and possibly automated, 'options' in generating useful and possibly non-obvious queries would be attractive. Data mining and machine learning techniques can be employed, through an inductive process, to discover automatically "queries" from a given dataset. More specifically, rule discovery and extraction algorithms can prove useful in "query formation". Examples of specific such algorithms are PART [7] and APRIORI [8].

1.1 Problem Statement

Validation of a complex Bayesian belief network, *i.e.*, one that has on the order of hundreds of variables, induced from a large dataset, like the National Crime Victimization Survey (NCVS), is a highly challenging task since it requires major investment of resources and domain expertise, while also being labor-intensive. The data mining and knowledge discovery algorithms are poised to offer a certain degree of relief from this challenge, and hence can be leveraged to automate segments of the overall process of query formation for validation. A machine learning or data mining algorithm can be leveraged to mine for rules in a dataset from which the Bayesian belief network model was induced, wherein these rules can be formulated as queries for validation purposes. The proposed study envisions processing a large and complex dataset through a rule-generation algorithm 1) to discover embedded knowledge in the form of if-then rules, and subsequently 2) to identify, through expert involvement, a subset of "interesting" and "significant" rules that can be formulated as queries for validation of the Bayesian belief network model of the dataset.

The next section discusses and elaborates on validation of a Bayesian belief network (BBN) model of a dataset, automatic query generation through a specific knowledge discovery tool, the NCVS dataset leveraged for this study, and the development of a BBN model on the same dataset. The subsequent section will demonstrate application of the proposed methodology to discover rules in the data set, filtering of rules to identify an interesting and significant subset, mapping of chosen rules into queries, and demonstration of application of such queries for validation purposes on a specific BBN model of a real-life size dataset that has over 250 attributes and 200,000 data points, namely the National Crime Victimization Survey.

2. Background

This section discusses fundamental aspects of the problem being addressed. Elaborations on validating Bayesian belief networks when employed as probabilistic reasoning models, query formulation with the help of machine learning and data mining, the dataset used for the study, National Crime Victimization Survey (NCVS), and the development of the Bayesian belief network model of the dataset are presented.

2.1 The NCVS Dataset

The National Crime Victimization Survey (NCVS) [9-10], previously the National Crime Survey (NCS), has been collecting data on personal and household victimization through an ongoing survey of a nationally representative sample of residential addresses since 1973. The geographic coverage is 50 United States. The 'universe' is persons in the United States aged 12 and over in "core" counties within the top 40 National Crime Victimization Survey Metropolitan Statistical Areas (MSA). The sample used was a stratified multistage cluster sample. The NCVS MSA Incident data that was chosen for this study contains select household, person, and crime incident variables for persons who reported a violent crime within any of the core counties of the 40 largest MSAs from January 1979 through December 2004. Household, per-

son, and incident information for persons reporting non-violent crime are excluded from this file. The NCVS, which contains 216,203 instances and a total of 259 attributes, uses a labeling system for the attributes represented by letters and numbers. A typical attribute of interest is labeled by a five character (alpha-numeric) tag, e. g., V4529.

2.2 Bayesian Belief Network Model of NCVS Data

A Bayesian belief network (BBN) expresses a view of the joint probability distribution of a set of variables, given a collection of independence relationships. This means that a Bayesian belief network will correctly represent a joint probability distribution and simplify the computations if and only if the conditional independence assumptions hold. The task of determining a full joint distribution, in a brute-force fashion, is daunting. Such calculations are computationally expensive and in some instances impossible. In order to address this formidable computational challenge, Bayesian belief networks are built upon conditional independence assumptions that appear to hold in many domains of interest.

A Bayesian belief network enables the user to extract a posterior belief. All causal relationships and conditional probabilities are incorporated into the network and are accessible through an automated inference process. A once tedious and costly (in terms of computation) method of extracting posterior beliefs in a given domain is now space-efficient and time-efficient. It is also possible to make queries on any attribute of one's choosing as long as it is one of those included in the model. One can easily adjust the prior evidence in the same manner enabling him to effectively compare and contrast posterior probabilities of a given attribute based on prior knowledge. The introduction of such a method has increased the breadth and depth of statistical analysis exponentially.

The BBN creation process consists of multiple phases. Following any preprocessing needed on a given dataset, the learning or training phase starts, wherein appropriate structure learner and parameter learner algorithms need to be selected by means of empirical means [11-17]. Learning a Bayesian belief network is a two stage process: first learn a network structure and then learn the probability tables. There are various software tools, some in the public domain and open source, to accomplish the development of a BBN through induction from data. For instance, the open-source and public-domain software tool WEKA [7], a machine learning tool that facilitates empirical development of clustering, classification, and functional approximation algorithms, has been leveraged to develop a BBN from the NCVS dataset for the study reported herein.

The validation phase can best be managed through a software tool that can implement the "probabilistic inferencing" procedure applicable for Bayesian belief networks. Another open-source and public-domain software tool, the JavaBayes [18] was used for this purpose, which is able to import an already-built BBN model, and facilitate through its graphical user interface querying of any attribute for its posterior probability value among many other options. A BBN model developed in WEKA can easily be imported into the JavaBayes. Once imported, the JavaBayes allows the user to identify and enter the evidence, and query a posterior belief of any attribute.

In this study, the BayesNet tool of the WEKA has been used to induce a classifier with the "Victimization" attribute in the NCVS dataset as the class label [19]. The NCVS dataset has been split into training and test subsets with 66% and 33% ratios, respectively. Simulations were run for a variety of structure and parameter learning options. Results suggest that a number of BBN models performed exceptionally well as classifiers for the "Victimization" attribute in the NCVS dataset. All WEKA versions of the local hill climbers and local K2 search algorithms led to classification performances on the test subset with 98% or better accuracy. Since the classification accuracy rates were so close to each other, the value of parameter "number of parent nodes" became significant given that it directly relates to the approximation capability of the BBN to the full joint distribution. Accordingly, the BBN model generated through the local K2 algorithm with Bayes learning and four parent nodes (the command-line syntax is "Local K2-P4-N-S BAYES" in WEKA format) was selected as the final network. This model, which, upon request, can be obtained in BIF format from the authors, has been used exclusively in the validation experiments reported in the following sections.

2.3 Validation of Bayesian Belief Networks

Validation of a Bayesian belief network is a comprehensive process. Once the Bayesian belief network (BBN) is induced from the data and subsequently tuned by the domain experts, the next step is the testing for validation of the premise that the network faithfully represents the full joint probability distribution subject to conditional independence assumptions [5,20,21]. As part of the validation task, values computed by the BBN are compared with those supplied by the domain experts, statistical analysis, and the literature. Another distinct activity for validation entails querying any variable for its posterior distribution or posterior expectation, and to obtain an explanation for a subset of or all of the variables in the network. In that respect, knowledge discovery and data mining tools, in conjunction with the domain experts, are leveraged to formulate a set of so-called "interesting" and "significant" queries to pose to the BBN. Validating a BNN is no trivial task and necessitates ad hoc and empirical elements. More specifically, a comprehensive and

rigorous process of evaluation and validation of a BBN model entails the following:

1) Perform elicitation review that consists of reviewing the graph structure for the model, and reviewing and comparing probabilities with each other [22].

2) Carry out sensitivity analysis that measures the effect of one variable on another [3].

3) Implement validation using the data that entails analysis of predictive accuracy and expected value calculations.

4) Conduct case-based evaluations that may include the following: run the model on test cases, compare the model output with the expert judgment, and finally, compare the model predictions with the "ground truth" or accepted trends currently relied upon by experts in the domain of interest.

The case-based evaluations validation step is the most costly and challenging since it requires substantial human expertise. In particular, elicitation of expert judgment to be leveraged for the validation of the Bayesian belief network poses a serious obstacle since numerous test cases or "queries" must be generated and applied to the Bayesian belief network model. The expected values must be defined in advance by human experts to form a basis for comparison with those calculated by the network itself.

2.4 Query Formulation

Machine learning and data mining techniques may be leveraged to automatically discover "queries" for a given dataset. A query is the calculation of the posterior probabilities of any attribute or variable based upon the given prior evidence. When a user provides that a specific attribute is observed to have a (discrete) value, this 'evidence' may be used in calculating the posterior probability of a dependent variable. This is best understood by an example. Assume that the user makes a query for the posterior probability that a person will be a victim of burglary. This query is dependent upon the values observed for relevant attributes like the gender of the potential victim. If burglary is shown to be dependent upon the gender of the victim, then the prior observed value of male or female for the potential victim's gender will need to be supplied by the user in order to calculate the conditional probability of this incident. This is analogous to an if-then rule: such a rule is a candidate for a query. One rule could postulate that

"**If** *the gender of the victim is female* **Then** *the probability of burglary will be greater than* 0.60."

By having such a rule at one's disposal, the process of making valid and knowledgeable queries can be streamlined. One does not necessarily have to solely rely on an expert for help to formulate "interesting" and "significant" queries. A rule set may be generated using one of many knowledge discovery algorithms, which can be structured to produce a set of if-then rules. Machine learning and data mining techniques prove useful for discovering knowledge that can be modeled as a set of if-then rules. Among the viable algorithms, PART [23], C4.5 or C5 [24], and RIPPER [25] from machine learning, and APRIORI [8] and its derivatives from the data mining fields are prominent.

3. Automation of Query Generation

This section presents application of machine learning algorithms for knowledge discovery in the form of if-then rules on the NCVS dataset for the purpose of formulating queries to the Bayesian belief network model of the same dataset. Although data mining algorithms are also appropriate for knowledge extraction and subsequent automation of the query formulation process [26], their computational cost may quickly become prohibitive if care is not exercised. Decision tree or list based algorithms within the domain of machine learning are appealing in that they can generate a rule set for a given single attribute of interest often within reasonable spatio-temporal cost bounds. Accordingly, the machine learning algorithm PART is chosen for the rule discovery and extraction task given its desirable algorithmic and computational properties. The PART algorithm [23] combines two approaches, C4.5 [24] and RIPPER [25] in an attempt to avoid their respective disadvantages. The main steps for validation of a Bayesian belief net model of data through automated query generation are shown in **Figure 1**.

The rule induction algorithm PART is applied to the NCVS dataset in order to extract a set of rules. The same rules are leveraged, following further processing by domain experts, as queries to the BBN model of the NCVS

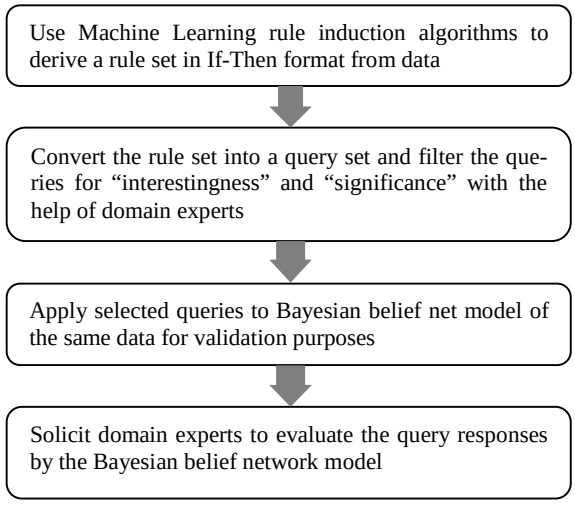

Figure 1. Generic overview of steps for Bayesian belief net validation through automated query generations

dataset for validation purposes. Initially, a subset of rules is labeled as "interesting" and "significant" by the domain experts, wherein "interesting" is a subjective labeling by a particular domain expert based upon the relationship of the evidence and the resultant projected probability of the THEN consequent variable. Next, these rules are formulated as queries and evidence associated with each query supplied to the BBN model on JavaBayes. Posterior probability calculations performed by the JavaBayes reasoning or inferencing engine for the attribute(s) of interest, which can be any subset from the list, are compared to expected values. This is done to infer if, in fact, the BBN model approximates reasonably well the joint probability distribution for the set of attributes entailed by the NCVS dataset.

3.1 PART Algorithm and Rules on NCVS Data

Rules that are derived from a dataset through a machine learning algorithm like PART expose the relationship between a subset of attributes and a single attribute of interest (or the class label), *i.e.* in this case the class label is designated as the "Victimization" due to its significance in the domain. Any attribute can be designated as the class label and would require a separate run of the PART algorithm to generate the set of rules whose consequents are the class label. Through the PART algorithm, the knowledge entailed by the dataset is captured into a framework with a set of if-then rules. Specifically, the format for a rule complies with the following: IF *premise* THEN *consequent*, where the premise is a statement of the form of a logical conjunction of a subset of attribute-value pairs, and the consequent represents a certain type of victimization. We have used the WEKA implementation of the PART algorithm throughout this study. Available options for the PART as implemented in the WEKA package and their associated default settings are shown in **Table 1**.

The NCVS Incident dataset was preprocessed prior to the rule induction step: the attribute count was reduced from 259 to 225 through removal of those that were not deemed to be relevant for the study. The attributes in the NCVS Incident dataset are represented, with a few exceptions, by a label that has four numeric characters preceded by the letter "V". The PART algorithm was applied to the NCVS dataset with default parameter values and the V4529 (Victimization) as the class attribute. Values for the V4529 attribute are shown in **Table 2**. The algorithm was trained on a 66%-33% training-testing split of the NCVS dataset, and generated a list of 176 rules [27]. The rules output are in the traditional IF-THEN format, where the premise is the logical conjunction of a set of attribute-value pairs (*i.e.*, evidence) followed by the consequent which is a specific value of the class attribute. **Table 3** illustrates one of the rules discovered by the PART algorithm on the NCVS data and its interpretation.

Table 1. Parameter options and default values for the WEKA PART algorithm.

PART Option	Explanation	Default Values
-C number	Confidence threshold for pruning	0.25
-M number	Minimum number of instances per leaf	2
-R	Use reduced error pruning	False
-N number	Number of folds for reduced error pruning	3
-B	Use binary splits for nominal attributes	False
-U	Generate unpruned decision list	False
-Q <seed>	Seed for random data shuffling	1

Table 2. Values for the NCVS attribute V4529

V4529 Label	Description of Values for "Victimization" Attribute V4529
x60	Completed/Attempted rape
x61	Sexual attack/assault/serious assault
x62	Attempted/completed robbery with injury from serious assault
x63	Attempted/completed robbery with injury from minor assault
x64	Attempted/completed robbery without injury
x65	Attempted/completed aggravated assault
x66	Threatened assault with weapon
x67	Simple assault completed with injury
x68	Assault without weapon without injury
x69	Verbal threat of rape/sexual assault
x70	Verbal threat of assault
x71	Attempted/Completed purse snatching and pocket picking
x72	Burglary
x73	Attempted forcible entry
x74	Attempted/completed motor vehicle theft
x75	Attempted/completed theft

3.2 Query Formulation Based on PART Rules

The process of query formulation using the PART rules and posing the queries to the BBN model entails human expert involvement and is the focus of the discussion in this section. A PART rule, which is captured through the "IF-*premise*-THEN-*consequent*" framework, readily lends itself to the query formation: the premise becomes the prior evidence for a query, where posterior probability value calculation is desired for the rule consequent. Such queries may be employed to validate, among other uses,

Table 3. A sample rule generated by the PART algorithm and its interpretation

PART Rule	Interpretation
V4113 = 0 & V4094 = 0 & V4119 = 0 & V4117 = 0 & V4118 = 0 & V4096 = 9:67	If the victim · did not receive injuries from an attempted rape (V4113 = 0), and · was not attacked in the form of rape (V4094 = 0), and · was not knocked unconscious (V4119 = 0), and · did not have broken bones or teeth as a result of incident (V4117 = 0), and · did not sustain any internal injuries (V4118 = 0), and · could not answer if (s)he was or was not a victim of sexual assault (V4096 = 9), Then · there is a high probability that this person will be a victim of "Simple Assault Completed with Injury" (V4529 = x67)

the Bayesian belief network model of the full joint probability distribution of the 225 attributes in the NCVS dataset. The list of 176 rules generated by the PART algorithm was manually processed by domain experts, Gabrielle Davis [28] and Michael Riesen [27], to identify those that are interesting and significant for query formation to serve as the validation set through the domain specialist's somewhat subjective perspective. The list of 49 rules identified accordingly to be leveraged as queries to the BBN model of the NCVS dataset are listed in [27].

Conversion of PART rules to queries and posing resulting queries to the JavaBayes realization of the BBN model is a straightforward process and will be illustrated next. The middle column in **Table 4** displays (in JavaBayes format) the posterior probability for the victimization attribute V4529 with no prior evidence observed before any query is posed as provided by the BBN model. One of the simple rules generated by the PART that will be used as an example query is shown in **Table 4**. The premise part of the rule, *i.e.*, V4127 = 2 AND V4095 = 1, is considered as prior evidence and supplied to the BBN model as such. Next, the JavaBayes is asked to perform "reasoning" or "inference" using the supplied prior evidence through the BBN model of the NCVS data. Once the inferencing calculations are complete, the updated posterior probabilities for all discrete values of the victimization attribute are as shown in the rightmost column in **Table 4**. As an example, the probability value for the *x*60 value of the victimization attribute is now 0.612, a marked increase compared to the no-evidence case. Translating the NCVS notation of the above comparison, this rule indicates that when a victim is attacked in such a way that the victim perceived the incident as an attempted rape (V4095 = 1) and the victim was not injured to the extent that the victim received any medical care, including self treatment (V4127 = 2), there is a 61% chance that this victim would be a victim of a completed rape or attempted rape (V4529 = *x*60).

Next, another and relatively more complex rule generated by the PART algorithm as shown in **Table 5** was presented as a query to the BBN model on JavaBayes. In Table 6, the process of supplying the evidence as provided from this PART rule is shown. First, the prior evidence that the victim suffered no injuries that are related to attempted rape (V4113 = 0) is supplied. Then, further prior evidence is supplied through V4052 = 0, meaning that the offender did not use a rifle, shotgun or any other gun different from a handgun. More prior evidence is added in the form of V4050 = 3, indicating that there was a weapon used, but the specific type is not applicable as reported in the NCVS. In the final step, V4241 = 1 as prior evidence is provided. However, with this addition of V4241 = 1 the JavaBayes running in the Java Runtime Environment generated an OutOfMemory exception, although the heap size was set to 3.5 GB. Nevertheless, for each of the reportable cases, the corresponding posterior probability table for the NCVS Victimization attribute V4529 is displayed. As shown in **Table 6**, inclusion of each further evidence has a direct affect on the posterior probability of the consequent (*i.e.*, the so-called "Then" part of a rule), which can be observed through the value of *x*65 discrete label for the class attribute V4529.

Table 4. A PART rule (V4127 = 2 & V4095 = 1: 60), associated JavaBayes query, and updated posterior probability values for V4529 with increasing evidence

Conditional Probabilities of V4529 Labels	Posterior Probabilities for V4529 with No Evidence	Posterier Probabilities with Evidence due to V4127 = 2 & V4095 = 1
p(x60|evidence)	0.004	0.612
p(x61|evidence)	0.001	0.005
p(x62|evidence)	0.005	0.006
p(x63|evidence)	0.005	0.009
p(x64|evidence)	0.025	0.003
p(x65|evidence)	0.036	0.003
p(x66|evidence)	0.006	0.069
p(x67|evidence)	0.022	0.008
p(x68|evidence)	0.055	0.003
p(x69|evidence)	0.000	0.007
p(x70|evidence)	0.019	0.227
p(x71|evidence)	0.018	0.010
p(x72|evidence)	0.113	0.007
p(x73|evidence)	0.032	0.009
p(x74|evidence)	0.053	0.008
p(x75|evidence)	0.598	0.008

Table 5. PART rule and associated JavaBayes query

PART Rule	Corresponding JavaBayes Query Syntax
V4113 = 0 AND V4052 = 0 AND V4050 = 3 AND V4241 = 1:65	Posterior distribution: probability ("V4529"|V4113 = 0, V4052 = 0, V4050 = 3, V4241 = 1)

Table 6. Posterior probabilities for the Victimization attribute V4529 with progressively increasing prior evidence (fraction truncated beyond third significant digit)

	Posterior Distributions		
V4529 Values	*probability* (*V*4529| *V*4113 = 0)	*probability* (*V*4529| *V*4113 = 0, *V*4052 = 0)	*probability* (*V*4529| *V*4113 = 0, *V*4052 = 0, *V*4050 = 3)
p(x60|evidence)	0.032	0.023	0.028
p(x61|evidence)	0.004	0.005	0.003
p(x62|evidence)	0.064	0.195	0.210
p(x63|evidence)	0.066	0.003	0.002
p(x64|evidence)	0.083	0.073	0.086
p(x65|evidence)	0.206	0.624	0.630
p(x66|evidence)	0.010	0.024	0.010
p(x67|evidence)	0.259	0.003	0.002
p(x68|evidence)	0.245	0.001	0.001
p(x69|evidence)	0.000	0.000	0.000
p(x70|evidence)	0.020	0.037	0.021
p(x71|evidence)	0.000	0.001	0.000
p(x72|evidence)	0.000	0.000	0.000
p(x73|evidence)	0.000	0.000	0.000
p(x74|evidence)	0.000	0.000	0.000
p(x75|evidence)	0.001	0.000	0.000

3.3 Validation of NCVS BBN Model through PART-Induced Queries

Each of 49 rules that were identified as "interesting" and "significant" by the domain experts was carefully considered as a test query. In light of the memory limitation encountered earlier, original rules had to be altered in order for the system to be able compute the posterior probabilities within the memory constraints of the system available. Accordingly, some of the rules were eliminated due to memory limitations: a total of 22 rules were selected, revised and included in the query list. **Table 7** shows a revised version of the rules supplied by the PART algorithm, which were computable and hence was applied as queries to the BBN model of the NCVS data. The attributes or evidence variables in each rule was ranked by domain experts [28-29], in order of interest (*i.e.* importance to study of the domain). The domain experts were able to classify two general groups of "interesting" and "significant" rules: 1) rules listing IF premises that produced an unexpected result; and 2) rules that were in direct alignment with the accepted standards in the domain.

Some attributes that are originally appearing in a specific rule and were ranked low by the experts were excluded from the corresponding query due to memory constraints. As a result of exclusion of certain attributes-value pairs from many of the 22 rules used as query, it is expected that the consequent attribute value is likely to be affected and possibly change from the value as indicated by the original rule induced by the PART rule discovery algorithm. Each revised rule in **Table 7** is indicated with an (R) next to the number of the rule.

The posterior probabilities of each rule in **Table 7** upon being posed as a query and as computed by the JavaBayes are displayed in **Table 8**, where only significant probability values are denoted for the sake of presentation clarity. **Table 9** represents the rules recovered from computed probabilities in **Table 8** to comparatively demonstrate the differences between the revised rules in **Table 7** and those computed by the BBN model of the NCVS data in **Table 9**. In formulating rules in **Table 9**, any consequent attribute value that has a comparatively significant probability value was included. Due to revision of the original rules induced from the NCVS data, there are differences between the consequents of rules in **Tables 7** and **9**.

Although there are discrepancies between the consequents of the rules in **Tables 7** and **9**, knowledge exposed by the PART rules is still present to a large degree. The "x75" represents the crime of attempted or completed theft and is a dominant value for the victimization attribute. With no evidence being presented, "x75" will represent nearly 60% of all crimes reported in the NCVS. Interestingly, the PART rules have extracted a second layer of usable information. The revised rules are not necessarily "incorrect" but are showing how a particular set of values can drastically affect the outcome of the victimization attribute. For example, rule 10 in unrevised form provides that the victimization attribute should have a large value for "x71". As noted in Tables 8 and 9, "x71" is not the dominant value for the revised rule 10. However, the change in posterior probability for the variable "x71" from 1.8% to 18% is nevertheless noteworthy. Where the rules generated by the PART algorithm are queried exactly as they appear, the consequents of the rule hold true as the dominant variable. Since certain queries fail due to memory error, rules had to be revised to demonstrate at least a portion of the knowledge extracted by the original PART-induced rules.

Table 7. Revised query list based on PART rules

Rule No	If	Then V4529 =	Rule No	If	Then V4529 =
1 (R)	V4065 = 1 & V4026 = 9 & V3018 = 1 & V3024 = 2	75	12 (R)	V4322 = 9 & V4065 = 1 & V4024 = 7 &	71
2 (R)	V4052 = 0 & V4083 = 9 & V4094 = 0 & V4095 = 0 & V4024 = 7	65	13 (R)	V4322 = 9 & V4065 = 1 & V4307 = 0 & V4024 = 8	71
3 (R)	V4052 = 0 & V4112 = 0 & V4113 = 0 & V4095 = 0 & V4094 = 0 & V4024 = 1	65	14	V4322 = 9 & V4065 = 1 & V4285 = 9 & V4307 = 0 & V4024 = 7 & MSACC = 35	71
4 (R)	V4052 = 0 & V4094 = 0 & V4095 = 0 & V4111 = 0 & V4024 = 2	65	15 (R)	V4322 = 9 & V4065 = 1 & V4024 = 3	71
5 (R)	V4322 = 9 & V4065 = 1 & V4024 = 5	71	16 (R)	V3024 = 2 & V3020 = 23 & V2045 = 1	71
6 (R)	V4322 = 9 & V4065 = 1 & V4024 = 7 & V3018 = 2 & MSACC = 17	71	23	V4073 = 0 & V4029 = 9 & V3018 = 2 & V4152 = 9 & V2045 = 2 & V3019 = 2	75
7 (R)	V4322 = 9 & V4065 = 1 & V4024 = 7 & V3018 = 2 & MSACC = 26	71	45 (R)	V4065 = 1 & V4029 = 9 & V3018 = 2 &	75
8 (R)	V4322 = 9 & V4065 = 1 & V4024 = 2	71	46 (R)	V3020 = 8	71
9 (R)	V4322 = 9 & V4065 = 1 & V4024 = 7 & MSACC = 4	71	47 (R)	V3020 = 24 & V3014 = 3	75
10 (R)	V4322 = 9 & V4065 = 1 & V4024 = 7 & V3015 = 5	71	48 (R)	V4113 = 0 & V4052 = 0 & V4050 = 3 &	65
11 (R)	V4322 = 9 & V4128 = 1 & V4094 = 0 & V4095 = 0 & V4052 = 0 & V4051 = 0 & V4289 = 2 &	65	35	V4322 = 9 & V4052 = 0 & V4081 = 9 & V4095 = 0 & V4094 = 0 & V4096 = 9 & V4036 = 9 & V4024 = 5	65

The query results for revised PART rules were reviewed by two domain experts [28,29]. In the majority of the cases, both experts found the predicted posterior probabilities to be reasonable and in accord with the current statistical trends provided by conventional means. As an example, the Bureau of Justice Statistics (BJS) provides periodic statistical reports [9]. BJS reported that, based upon violent crimes statistics from 1973-2005, beginning with the 25-34 age category, the rate at which persons were victims of violent crimes declined significantly as the age category increased [30]. The BJS also reports that in general, males experienced higher victimization rates than females for all types of violent crime except rape/sexual assault [9]. Where the generated rules included attributes (e.g. V3014 (Age), V3018 (Gender), and V4024 (location of incident)) that were consistent with known and generally accepted trends, the experts were not surprised with the values predicted and agreed that the posterior probabilities based upon each set of the evidence attributes were not in the extremes, based upon current publications in the field. The values were not unexpectedly high and thus did not trigger a shocking response. Conversely, the posterior values were not inordinately low compared to expected results, and thus the validity of the predicted value was not drawn into question.

Rules 11, 35 and 48 were highlighted by the experts as the strongest rules, having the most sensible values for posterior prediction as compared to the generally accepted statistical values presented in currently available publications and studies. In particular, the experts easily identified a known relationship or correlation between the IF premise and consequent for each of the rules 11, 35, and 48. In each of these three strongest rules, experts found the prior evidence values clearly set the stage for the associated posterior victimization predictions. Overall, both experts indicated that the responses computed by the BBN model of the NCVS data to all queries posed were expected and reasonable in generality, suggesting that the model is realistic, and accordingly is a good approximation to the joint probability distribution.

As an exception to the generally positive feedback, rule 10 was found to be somewhat extraordinary. Rule 10 included the attribute that the victim was never married (V3015 = 5). A value of 5 for V3015 shows a distinct increase for the probability of a purse snatching or pick-pocketing. Domain experts were surprised to find that this evidence value would have such an impact on the posterior probability of pick pocketing. Although the posterior prediction was not necessarily discounted, experts were skeptical, outside a more thorough explanation of the increased victimization. However, the skepticism did not detract from the intriguing prospect that the generated rule might have exposed "new" knowledge. As the experts reviewed the list of rules, the inclusion of certain "unusual" or unexpected attributes similar to the attribute uncovered by rule 10 stimulated the most feedback from the domain experts. The experts were interested in further investigation of the "new" and "unusual"

Table 8. Query results as probability values for revised PART rules in Table 7 (only highest probability values are shown and fractions are truncated beyond the second significant digit)

Rule No	x61	x62	x64	x65	x66	x67	x68	x69	x70	x71	x72	x73	x74	x75
1									0.04	0.20			0.03	0.68
2		0.26	.09	**0.61**	0.01									
3		0.22	.06	**0.67**										
4		0.23		**0.74**										
5					0.05				0.13	0.06		0.02	**0.36**	**0.32**
6					0.04				0.09	**0.28**			0.12	**0.41**
7					0.03			0.01	0.08	**0.36**			0.09	**0.38**
8					0.02				0.14	0.02		0.01	**0.30**	**0.46**
9					0.04				0.15	**0.21**			0.12	**0.41**
10					0.07				0.17	0.18		0.01	0.07	**0.43**
11				**0.98**	0.01									
12					0.05				0.15	0.19			0.09	**0.44**
13					0.01					0.12			0.07	**0.69**
14					0.01				0.03	**0.35**			0.08	**0.49**
15	0.01				0.07				0.16	0.04		0.02	0.14	**0.47**
16			0.02	**0.03**							0.11	0.03	0.05	**0.59**
23									0.01	**0.20**	0.11		0.02	**0.63**
35		0.09	0.06	**0.78**	0.03									
45									0.02	0.20	0.10		0.03	**0.62**
46			0.03	**0.04**		0.03	0.07				0.08		0.04	**0.59**
47							0.05				0.10	0.03	0.05	**0.60**
48		0.21	0.08	**0.63**					0.02					

combination of attribute-value pairs presented in generated rules, stating that the rules could provide a starting point for further research of factors that may not have been fully developed with conventional methods.

The implications of using a rule generating algorithm such as the PART to essentially generate queries are potentially profound. Limitations associated with user bias and limited domain knowledge may impede the self-generation of useful and interesting queries. Using PART as an automatic query generation tool could potentially uncover a not-so-obvious relationship between prior evidence and the resulting posterior probability of another attribute. Applying this principle to the NCVS data, the practical significance means uncovering the specific attributes of a victim or circumstance that makes them more or less probable to be a victim of a specific crime. As an example of practical implementation within the context of criminal justice, by identifying these relationships that have the greatest impact on posterior probability, resources can be channeled into areas that would be most effective in combating violent crime.

Domain experts indicated that automatic query generation using the PART algorithm or an equivalent would be helpful in not only discovering any hidden or novel relationships between attributes, but more practically as a method to reinforce trends and relationships already relied upon in the field. A second group of domain experts[1] were independently interviewed and asked to provide a list of self-generated queries that would be of personal interest. None of the second group was able to provide a list of more than three potential queries. The second group was then presented with the automatically generated queries. All experts in the second group found that

[1]Six Professors at the University of Toledo College of Law

Table 9. Rules reconstructed from probability values in Table 8 (only modified rules are shown)

Rule No	If	Then V4529 =	Rule No	If	Then V4529 =
5 (R)	V4322 = 9 & V4065 = 1 & V4024 = 5	74 & 75	14	V4322 = 9 & V4065 = 1 & V4285 = 9 & V4307 = 0 & V4024 = 7 & MSACC = 35	71 & 75
6 (R)	V4322 = 9 & V4065 = 1 & V4024 = 7 & V3018 = 2 & MSACC = 17	71 & 75	15 (R)	V4322 = 9 & V4065 = 1 & V4024 = 3	704 & 74 & 75
7 (R)	V4322 = 9 & V4065 = 1 & V4024 = 7 & V3018 = 2 & MSACC = 26	71 & 75	16 (R)	V3024 = 2 & V3020 = 23 & V2045 = 1	72 & 75
8 (R)	V4322 = 9 & V4065 = 1 & V4024 = 2	74 & 75	23	V4073 = 0 & V4029 = 9 & V3018 = 2 & V4152 = 9 & V2045 = 2 & V3019 = 2	7 1 & 75
9 (R)	V4322 = 9 & V4065 = 1 & V4024 = 7 & MSACC = 4	71 & 75	35	V4322 = 9 & V4052 = 0 & V4081 = 9 & V4095 = 0 & V4094 = 0 & V4096 = 9 & V4036 = 9 & V4024 = 5	62 & 65
10 (R)	V4322 = 9 & V4065 = 1 & V4024 = 7 & V3015 = 5	70 & 71 & 75	45 (R)	V4065 = 1 & V4029 = 9 & V3018 = 2 &	71 & 75
12 (R)	V4322 = 9 & V4065 = 1 & V4024 = 7 &	70 & 71 & 75	46 (R)	V3020 = 8	75
13 (R)	V4322 = 9 & V4065 = 1 & V4307 = 0 & V4024 = 8	71 & 75			

the collection of automatically generated queries was relatively easy to review compared to the alternative of postulating the-defined list of rules and queries.

Each of the experts in the second group agreed that it is sometimes difficult to consider the impact of a particular variable, especially if the particular variable is not one that has been extensively researched using other known techniques. In this way, the automatic rule generation may also be used as a reliable method to test prior hypotheses. Each member of the second group also agreed that an automatically generated list of rules provided a catalyst to the generation of user-defined rules and queries. At a minimum the relationships of the attributes presented in the generated rules caused members in the second group to reflect upon their own conception of trends in victimization, which ultimately resulted in a wholesale request for more information on the resultant effect of certain unexpected attributes on the posterior probability of victimization.

4. Conclusions

This paper presented an approach to address the acquisition bottleneck problem in generating human expert-formulated queries for validation of a Bayesian belief network model. A machine learning based approach for rule discovery from a dataset to serve as potential queries was proposed. The proposed technique employs machine learning (and potentially data mining) algorithms to generate a set of classification or association rules that can be converted into corresponding queries with minimal human intervention and processing in the form of filtering for interestingness and significance by domain experts. The application and utility of proposed methodology for semi-automated query formulation based on rule discovery was demonstrated on validation of a Bayesian belief network model of a real life size dataset from the domain of criminal justice.

REFERENCES

[1] D. Heckerman, "Bayesian Networks for Data Mining," *Data Mining and Knowledge Discovery*, Vol. 1, No. 1, 1997, pp. 79-119.

[2] K. B. Laskey and S. M. Mahoney, "Network Engineering for Agile Belief Network Models," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 12, No. 4, 2000, pp. 487-498.

[3] K. B. Laskey, "Sensitivity Analysis for Probability Assessments in Bayesian Networks," *Proceedings of the Ninth Annual Conference on Uncertainty in Artificial Intelligence*, Washington, D.C., 1993, pp. 136-142.

[4] M. Pradham, G. Provan, B. Middleton and M. Henrion, "Knowledge Engineering for Large Belief Networks," *Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence*, Seattle, Washington, 1994, pp. 484-490.

[5] O. Woodberry, A. E. Nicholson and C. Pollino, "Parameterising Bayesian Networks," In: G. I. Webb and X. Yu Eds., *Lecture Notes in Artificial Intelligence*, Springer-Verlag, Berlin, Vol. 3339, 2004, pp. 1101-1107.

[6] S. Monti and G. Carenini, "Dealing with the Expert Inconsistency in Probability Elicitation," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 12, No. 4, 2000, pp. 499-508.

[7] H. Witten and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques," 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[8] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," *Proceedings of the* 20*th International Conference on Very Large Data Bases*, Santiago, 1994,

pp. 487-499.

[9] US Department of Justice, Bureau of Justice Statistics. National Crime Victimization Survey: Msa Data, 1979-2004. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2007-01-15. http://www.icpsr.umich. edu/cocoon/NACJD/STUDY/04576.xml

[10] T. C. Hart and C. Rennison, Bureau of Justice Statistics, "Special Report", March 2003, NCJ 195710. http://www.ojp.usdoj.gov/bjs/abstract/rcp00.html

[11] R. Blanco, I. Inza and P. Larrañaga, "Learning Bayesian Networks in the Space of Structures by Estimation of Distribution Algorithms," *International Journal of Intelligent Systems*, Vol. 18, No. 1, 2003, pp. 205-220.

[12] R. Bouckaert, "Belief Networks Construction Using the Minimum Description Length Principle," *Lecture Notes in Computer Science*, Springer-Verlag, Berlin, Vol. 747, 1993, pp. 41-48.

[13] L. M. de Campos, J. M. Fernández-Luna and J. M. Puerta, "An Iterated Local Search Algorithm for Learning Bayesian Networks with Restarts Based on Conditional Independence Tests" *International Journal of Intelligent Systems*, Vol. 18, No. 2, 2003, pp. 221-235.

[14] J. Cheng, R. Greiner, J. Kelly, D. A. Bell and W. Liu, "Learning Bayesian Networks from Data: An Information—Theory Based Approach," *Artificial Intelligence*, Vol. 137, No. 1-2, 2002, pp. 43-90.

[15] D. Heckerman, D. Geiger and D. M. Chickering, "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data," *Machine Learning*, Vol. 20, No. 3, 1995, pp. 197-243.

[16] T. V. Allen and R. Greiner, "Model Selection Criteria for Learning Belief Nets: An Empirical Comparison," *Proceedings of International Conference on Machine Learning*, Stanford, 2000, pp. 1047-1054.

[17] Y. Guo and R. Greiner, "Discriminative Model Selection for Belief Net Structures," *Proceedings of the Twentieth National Conference on Artificial Intelligence*, Pittsburgh, 2005, pp. 770-776.

[18] F. G. Cozman**, "**JavaBayes Software Package," University of São Paulo, Politécnica**,** cited 2006. http://www.cs.cmu.edu/~fgcozman/home.html

[19] R. Bouckaert, "Bayesian Network Classifiers in Weka," Technical Report, Department of Computer Science, Waikato University, Hamilton, 2005.

[20] M. J. Druzdzel and L. C. van der Gaag, "Building probabilistic Networks: Where do the Numbers Come from?" *IEEE Transactions on Knowledge and Data Engineering*, Vol. 12, No. 4, 2000, pp. 481-486.

[21] T. Boneh, "Visualisation of Structural Dependencies for Bayesian Network Knowledge Engineering," Masters Thesis, University of Melbourne, Melbourne, 2002.

[22] M. J. Druzdzel and L. C. van der Gaag, "Elicitation of Probabilities for belief Networks: Combining Qualitative and Quantitative Information," *Proceedings of the Tenth Annual Conference on Uncertainty in AI*, Seattle, 1995, pp. 141-148.

[23] H. Witten and E. Frank, "Generating Accurate Rule Sets without Global Optimization," *Proceedings of the Fifteenth International Conference on Machine Learning*, Madison, 1998, pp. 144-151.

[24] J. R. Quinlan, "C4.5: Programs for Machine Learning," Morgan Kaufmann Publishers, San Mateo, 1993.

[25] W. W. Cohen, "Fast Effective Rule Induction," *Proceedings of the* 12*th International Conference on Machine Learning*, Lake Tahoe, 1995, pp. 115-123.

[26] J. Hipp, U. Guntzer and G. Nakaeizadeh, "Algorithms for Association Rule Mining—A General Survey and Comparison," *ACM SIGKDD Explorations*, Vol. 2, No. 1, 2000, pp. 58-64.

[27] M. Riesen, "Development of a Bayesian Belief Network Model of NCVS Data as a Generic Query Tool," Masters Project, Engineering, University of Toledo, Toledo, 2007.

[28] G. Davis, Private communications, College of Law, University of Toledo, Toledo, 2008.

[29] P. Ventura, "Private Communications, Criminal Justice," University of Toledo, Toledo, 2008.

[30] S. M. Catalano, Crime Victimization 2005, NCJ 214644. http://www.ojp.usdoj.gov/bjs/abstract/cv05.html

