A Bayesian Belief Network Classifier for Predicting Victimization in National Crime Victimization Survey

Michael Riesen[1] **and Gursel Serpen**[2]
[1]College of Law, University of Toledo, Toledo, OH 43606 USA
[2]Electrical Engineering and Computer Science
College of Engineering, University of Toledo, Toledo, OH, 43606 USA

Abstract *- This paper presents the development of a Bayes net classifier for prediction of a victimization attribute value for the National Crime Victimization Survey dataset. The National Crime Victimization Survey dataset has over 250 attributes and 216,000 data points, and as such poses a large-scale problem context for classifier development. The classifier was developed using the Weka machine learning software workbench. A set of structural and parameter learning algorithms for the Bayesian belief network were employed in a development effort while ensuring that the computational complexity in both time and space remained within affordable bounds. A number of structural learning algorithms, including local versions of hill-climbing and K2, provided a classification performance of 99% on the testing data. Simulation results indicate that it is feasible to develop a successful Bayesian belief network classifier for the victimization attribute of the National Crime Victimization Survey data.*

Keywords: Bayesian belief net, classifier, prediction, national crime victimization survey, victimization

1 Introduction

A dataset can serve as a valuable source of information and knowledge about a domain. One way to leverage the embedded knowledge of such a dataset is to develop predictors on certain attributes of interest through empirical means. Mainly, statistics, and to an ever increasing degree, machine learning approaches, can accomplish the task of predictor development equally well.

The National Crime Victimization Survey (NCVS) is a rich source of knowledge within the criminal justice domain [US DOJ, 2007]. Automated software tools that can extract this knowledge, offer significant benefits to the criminal justice community as well as the public at large. In that context, the inductive learning approaches from the machine learning domain are appropriate for empirical development of predictors for a given attribute in the NCVS dataset. For example, Bayesian Belief networks provide a means for such an empirical development of predictors. Bayesian belief networks are probabilistic modeling tools and can approximate the posterior probability distribution of any chosen attribute in the domain. In fact, Bayesian belief networks were employed in a variety of ways to develop probabilistic models in the criminal justice domain [Baumgartner, 2005; Huygen, 2002; Garbolina, 2002; Leucari, 2006; Muecke, 2007; Pordoe, 2006; and Strnad, 2006]. In light of several sources of uncertainty associated with the NCVS, a probabilistic modeling tool like the Bayesian belief network (BBN) is a good option for classifier development on the same dataset. This paper proposes leveraging the Bayesian belief network as an empirical classifier model for predicting the value of the "victimization" attribute in the NCVS data.

The following sections present the details of the work accomplished. The NCVS dataset and the preprocessing applied is discussed next. This is followed by a presentation on the issues related to Bayesian belief network classifier development. The simulation study and its results are reported subsequently. Finally, conclusions follow.

2 Data and Pre-Processing

The National Crime Victimization Survey (NCVS) series, previously called the National Crime Survey (NCS), has been collecting data on personal and household victimization since 1973 [US DOJ, 2007]. An ongoing survey of a nationally representative sample of residential addresses, the NCVS is the primary source of information on the characteristics of criminal victimization and on the number and types of crimes reported to law enforcement authorities. It provides the largest national forum for victims to describe the impact of crime and characteristics of violent offenders. Twice each year, data are obtained from a nationally representative sample of roughly 49,000 households comprising about 100,000 persons on the frequency, characteristics, and consequences of criminal victimization in the United States. The survey is administered by the Census Bureau (under the U.S. Department of Commerce) on behalf of the Bureau of Justice Statistics (under the U.S. Department of Justice).

In this particular study, we focus on the extract or supplement files created from the NCVS and NCS data series. Particularly two data files are of special interest, a weighted person-based file, and a weighted incident-based file, which contain the "core" counties within the top 40 National Crime Victimization Survey Metropolitan Statistical Areas (MSAs). Core counties within these MSAs are defined as those self-representing primary sampling units that are common to the MSA definitions determined by the Office of Management and Budget for the 1970-based, 1980-based, and 1990-based sample designs. Each MSA is comprised of only the core counties and not all counties within the MSA. The person-based file contains select household and person variables for all people in NCVS-interviewed households in the core counties of the 40 largest MSAs from January 1979 through December 2004. The 40 largest MSAs were determined based on the number of household interviews in an MSA. The incident-based file contains select household, person, and 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, person, and incident information for persons reporting non-violent crime are excluded from this file.

The NCVS MSA dataset [US DOJ, 2007] originally consisted of 259 discrete-valued attributes (variables) and 216,000 distinct instances of reporting. Due to redundancy and given the scope of our study, a total of 34 attributes were removed. Additionally, missing values for any of the attributes are potentially problematic for any dataset. In the case of the NCVS this concern was alleviated to a large degree since it was already preprocessed to address this issue. In the NCVS dataset, a numeric value of "9" was assigned for any missing value, which represented the so-called "out of universe" as elaborated in the following:

"… *Out of Universe (or INAP) is used in the codebook documentation to designate those items on the questionnaire that were not appropriate for certain respondents (based on information gathered throughout the interview) and therefore should not have been asked. For example, hospital tenure questions are not asked of victims who were not injured*…." [US DOJ, 2007]

The INAP value therefore alleviates any pre-processing concerns for missing or unusable values.

The Bayesian belief network implementation in the Weka software package requires that all discrete values for every attribute be a nominal value. Accordingly, the numeric values were converted into discrete nominal values. To make sure that all values were proper and included, we cross referenced the values with the codebook provided by the Inter-university Consortium for Political and Social Research (ICPSR).

3 Simulation Study

3.1 Learning Bayes Net

A brief discussion herein provides the main highlights of the major issues pertaining to developing a Bayesian belief network classifier through predominantly empirical analysis. Inducing a Bayesian belief network from data requires an appropriate structure-learning algorithm, a scoring function, and a parameter-learning algorithm to be identified (all three often through empirical means). The K2 algorithm is the choice of many researchers for Bayesian network structure learning [de Campos et. al., 2003 & 2000; Madden, 2003]. Other structure learning algorithms include naïve Bayes, conditional independence, hill climbing, simulated annealing, and tabu search among others [Van Allen and Greiner, 2000]. Prominent scoring functions include the so-called Bayes [Cooper et al., 1992], minimum description length (MDL) principle [Bouckert, 1993], Akaike's information criterion (AIC) [Akaike, 1974], and entropy [Cheng et al., 2002; Herskowitz, 1990]].

3.2 Weka and Bayes Net

The machine learning software workbench Weka [Witten and Frank, 2000] will be used for conducting the empirical analysis. The BayesNet classifier algorithm in Weka (Version 3.5.5) will be leveraged to develop the Bayesian Belief Network [Bouckert, 2005]. The options that must be addressed in Weka include the estimator that computes the conditional probability tables of the Bayes network, the searchAlgorithm that implements a user-selected structure learning algorithm, and the useADTree that facilitates savings in learning time at the expense of increased memory usage. The estimator will be set to SimpleEstimator with the default alpha value of 0.5, while the useADTree parameter will be set to false since the data set is poised to demand substantial memory due to its space complexity. The so-called searchAlgorithm option with Weka will be set to a number of choices (as elaborated upon below) in order to adequately explore the structure learning space.

The structure learning algorithms as implemented in Weka (through the so-called searchAlgorithm option) are presented in three groupings: local score based structure learning (i.e., minimum description length principle based), conditional independence based structure learning, and global score based structure learning (i.e., cross validation based). The local score based structure learning algorithms are desirable for computation cost savings purposes. We employed three of these algorithms, which included K2, hill climber, and tabu search. The set of local score based algorithms and associated option settings are presented in Tables 1 through 3. The conditional independence based structure learning option will also be explored and its realization within Weka, the CISearch, will be experimented with through the settings indicated as in

Table 4. Additionally, the naïve Bayes classifier algorithm (through default values) was also tested on the NCVS dataset to serve as a standard to compare against.

Table 1. Option Settings for the K2 Algorithm.

K2	Option Settings & Rationale
initAsNaiveBayes	false (an empty network is used as the initial network structure)
MarkovBlanketClassifier	false (this setting is left at its default value since not deemed to be critical for our purposes)
maxNrOfParents	3 or 4
scoreType	{Bayes, MDL, AIC, and Entropy}
Random Order	false

Table 2. Option Settings for the Hill Climber Algorithm.

Hill Climber	Option Settings & Rationale
InitAsNaiveBayes	false (an empty network is used as the initial network structure)
MarkovBlanketClassifier	false (this setting is left at its default value since not deemed to be critical for our purposes)
maxNrOfParents	3
scoreType	{Bayes and Entropy}
useArcReversal	true

Table 3. Option Settings for the Tabu Search Algorithm.

Tabu Search	Option Settings & Rationale
initAsNaiveBayes	false
markovBlanketClassifier	false
maxNrOfParents	1000
Runs	10
scoreType	Bayes
tabuList	5
useArcReversal	true

Table 4. Option Settings for the CISearch Algorithm.

CISearch	Option Settings & Rationale
markovBlanketClassifier	false
scoreType	Bayes

The set of options for the local score based search algorithms are initAsNaiveBayes, MarkovBlanketClassifier, maxNrOfParents, scoreType, useArcReversal, and randomOrder. The parameter initAsNaiveBayes has two settings. A value of true, which is the default, results in a naive Bayes network structure to be used as the initial network structure. On the other hand a false value will impose an empty network structure initially, i.e., the Bayes net has no arrows.

The markovBlanketClassifier (set to false by default), if set to true, leverages a heuristic: when the search space is traversed completely, this heuristic is used to validate that each of the attributes are in the Markov blanket of the classifier node. If a node that is not already in the Markov blanket (i.e., is a parent, child of sibling of the classifier node), an arrow is added. If the value of this parameter is set to false, then no action is taken.

The scoreType parameter is used to identify the score metric to be used. The set of available score metrics include Bayes, AIC, Entropy, and MDL. The maxNrOfParents parameter establishes an upper bound on the number of parents for each node in the network.

The randomOrder parameter has a default value of false, which implies that the order of the nodes in the dataset is used. If the randomOrder parameter is set to true, then the order of nodes in the network is randomly determined.

The parameter useArcReversal has a default value of false, and when set to true results in arc reversal operation to be performed during the search.

3.3 Simulation Results

The Bayesian belief network structure learning algorithms in Tables 1 through 4 and the naïve Bayes algorithm were trained and tested on the revised NCVS Incident dataset that had 225 attributes (approximately 10% of attributes were excluded due to irrelevance or redundancy) and 216,000 instances. The JavaHeap size was set to 3.5 GB for Weka, which was invoked in command-line mode to facilitate incremental loading of the input data file among other desirable aspects. The simulation platform is a Sun 4-Sparc-processor system with 8 GB shared RAM under Solaris™ 10 operating system.

The parameter that controls the number of parents (maxNrOfParents) appeared to be the most costly in terms of memory space and accordingly the most constraining. The value of this parameter had to be dramatically scaled down to three or four, the latter in some limited cases only, to be able to facilitate any models to be built at all: the JavaHeap size of 3.5 GB was simply inadequate for any model to be constructed for values greater than four for this parameter. More specifically, if we imposed no restrictions on the maximum number of parent nodes, no BBN model could be built with the local search algorithms. Through a series of tests and a reliance on previous literature [Livingston, 2005], we were able to build completed models using a maximum of three parent nodes. The computational cost of memory and time grew exponentially as the maximum number of parent nodes was increased beyond three. The local K2 search algorithm was the only model to be successfully built with more than three parent nodes, (using a maximum of four parent nodes). However, this might not have as dramatic impact on the approximation capacity of the BBN on the posterior

distribution for the class attribute "victimization" as elaborated in detail in [Livingston 2005].

It was also discovered that the choice of scoring function (i.e. MDL or BAYES) has a much greater effect on the sensitivity and precision of the resulting model than the choice of search method (i.e. K2 or hill-climbing) [Shaughnessy, 2005]. The choice of scoring function involves a tradeoff between sensitivity (learning many correct edges) and precision (learning only correct edges), and so the decision should depend on what an experimenter intends as the purpose of the network. In some other cases, we were unable to facilitate a particular search algorithm to complete even in light of the findings above for those factors impacting the space complexity.

It should also be noted that, due to the costly memory use of cross validation, building and testing a model on the NCVS dataset by means of cross validation proved to be not feasible. Instead, we implemented a 66%-33% split of data set, and trained the model on 66% of the full dataset while using the remaining 33% for testing. Highlights of successful runs that led to a complete BBN model are presented in Table 5.

Table 5. Successful BBN Classifier Models for Victimization Attribute and Associated Performance Profiles.

Search Algorithm (including options in Weka format)	Build Time (sec)	Test Time (sec)	% correct (Training)	% correct (Testing)
Naïve Bayes	9.89	77.16	76.46	76.41
CISearch (-S BAYES)	38.77	197.72	76.68	76.63
Local Hill Climber (-P 3 -N -S BAYES)	50649.06	256.42	99.28	99.22
Local Hill climber (-P 3 -N -S ENTROPY)	49812.78	301.30	99.61	98.77
Local K2 (-P 3 -N -S AIC)	17883.23	245.48	98.98	98.86
Local K2 (-P 3 -N -S BAYES)	17946.42	252.16	98.93	98.88
Local K2 (-P 3 -N -S ENTROPY)	21296.36	312.91	99.15	98.17
Local K2 (-P 3 -N -S MDL)	20286.85	224.92	97.13	97.08
Local K2 (-P 4 -N -S BAYES)	23616.75	342.25	99.21	99.19
Local Tabu Search (-R -N -U 10 -P 3 –S BAYES)	33336.49	192.83	59.87	59.76

3.4 Analysis

The naïve Bayes algorithm performance, which is 76% for the testing split of the NCVS data, serves as a standard against which performances of other algorithm can be compared. Results in Table 5 suggest that a number of BBN models performed exceptionally well as classifiers for the "victimization" attribute. In fact, all listed versions of the local hill climbers and local K2 search algorithms led to classification performances with 97% or better accuracy on the testing data. On the other hand, performances of CISearch and local tabu search are drastically low as to further suggest that the NCVS dataset does not necessarily pose an easy classification problem.

It is noted that the built time for the algorithms among the leading performers are on the same order and reasonable. The fundamental challenge in prototyping a classifier algorithm was the memory space cost and not necessarily the execution time cost. As an example, all versions of the simulated annealing algorithm tested on the NCVS dataset failed to complete the training due its high memory cost. Many versions of the Bayesian belief network with a value greater than four for the maximum number of parents parameter also failed to complete the training phase due to excessive memory cost given the computing platform available for this research study.

4 Conclusions

A Bayesian belief network classifier that predicts victimization in the National Crime Victimization Survey data has been successfully developed. A specific instance of this Bayes net classifier (local hill climber with a maximum of three parent nodes and entropy score metric) demonstrated a prediction accuracy of 99.22% on the testing data. The Bayesian belief net classifier offers an effective means to assess the likelihood of victimization. As such it is well positioned to serve as a prediction tool for victimization in the NCVS data for the criminal justice domain. This study can easily be extended to any other attribute of interest in the NCVS dataset for a similar classifier development using Bayesian belief networks.

5 References

[1] H. Akaike. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19:716–723, 1974.

[2] R. R. Bouckaert. Belief networks construction using the minimum description length principle. Lecture Notes in Computer Science, 747:41–48, 1993.

[3] Baumgartner, K.C.; Ferrari, S.; Salfati, C.G., 2005, "Bayesian Network Modeling of Offender Behavior for Criminal Profiling", Decision and Control, 2005 and 2005

European Control Conference. CDC-ECC apos;05. 44th IEEE Conference on Volume , Issue , 12-15 Dec. 2005 Page(s): 2702 - 2709

[4] Bouckaert, R, 2005, “Bayesian Network Classifiers in Weka”, Technical Report, Department of Computer Science, Waikato University, Hamilton, NZ 2005.

[5] 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, 18:221–235, 2003.

[6] L. M. de Campos and J. F. Huete. A new approach for learning belief networks using independence criteria. International Journal of Approximate Reasoning, 24:11–37, 2000.

[7] J. Cheng, R. Greiner, J. Kelly, D. A. Bell, and W. Liu. Learning Bayesian networks from data: an information-theory based approach. Artificial Intelligence, 137:43–90, 2002.

[8] G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. Machine Learning, 9:309–348, 1992.

[9] Garbolino, P. & Taroni, F., 2002, “Evaluation of scientific evidence using Bayesian networks”. Forensic Sci. Int. 125, 149–155 (2002).

[10]E. Herskovits and G. F. Cooper. Kutató: An entropy-driven system for the construction of probabilistic expert systems from databases. In Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence, pages 54–62, 1990.

[11]Huygen, P. E. (2002). “Use of Bayesian Belief Networks in Legal Reasoning”. In: Proc. of the 17th BILETA Ann. Conf. Amsterdam. Cf. also at http://www.bileta.ac.uk/02papers/ huygen.html. 2002

[12]Livingston, Gary, and Shaughnessy, Patrick (2005), “Evaluating the Causal Explanatory Value of Bayesian Network Structure Learning Algorithms”. Submitted to SIAM SDM 2006 (Society for Industrial and Applied Mathematics Conference on Data Mining).

[13]Leucari, Valentina, 2006, “Evidence Seminar: BAYESIAN NETWORKS FOR THE ANALYSIS OF EVIDENCE”, March 20th, 2006 Basement Lecture Theatre, 1-19 Torrington Place, London WC.

[14]Madden, M. (2003). “The performance of Bayesian network classifiers constructed using different techniques”. In Working notes of the ECML/PKDD-03 workshop on probabilistic graphical models for classification (pp. 59–70). 2003

[15]Muecke, Nial; Stranieri, Andrew, 2007, “An argument structure abstraction for Bayesian Belief Networks: just outcomes in on-line dispute resolution”, ACM International Conference Proceeding Series; Vol. 247 archive Proceedings of the fourth Asia-Pacific conference on Conceptual Modeling - Volume 67 Ballarat, Australia Pages: 35 – 40, 2007

[16]Pardoe, I. and R. R. Weidner (2006). “Sentencing convicted felons in the United States: a Bayesian analysis using multilevel covariates (with discussion)”. Journal of Statistical Planning and Inference, 136(4) 1433-1472

[17]Strnad, Jeff, 2007, “Should Legal Empiricists Go Bayesian?” Stanford Law and Economics Olin Working Paper, No. 342, 2007.

[18]Witten, Ian H. and Frank, Eibe (2005) "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[19]Van Allen, T., & Greiner, R. (2000). “Model selection criteria for learning belief nets: An empirical comparison”, In International Conference on Machine Learning (pp. 1047–1054).

[20]U.S. Dept. of Justice, Bureau of Justice Statistics. NATIONAL CRIME VICTIMIZATION SURVEY: MSA DATA, 1979-2004 [Computer file]. Conducted by U.S. Dept. of Commerce, Bureau of the Census. ICPSR04576-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producer and distributor], 2007-01-15.
http://www.icpsr.umich.edu/cocoon/NACJD/STUDY/04576.xml