RESEARCH REPORT SERIES
(*Statistics #2002-09)*

Bayesian Networks Representations, Generalized Imputation, and Synthetic Micro-data Satisfying Analytic Constraints

Yves Thibaudeau and William E. Winkler

Statistical Research Division
U.S. Bureau of the Census
Washington D.C. 20233

Report Issued: November 22, 2002

Bayesian Networks Representations, Generalized Imputation, and Synthetic Micro-data satisfying Analytic Constraints

Yves Thibaudeau[1] and William E. Winkler[1] 2002Oct24
{yves.thibaudeau,william.e.winkler}@census.gov

Abstract

This paper shows how Bayesian Networks can be used to create models for discrete data from contingency tables. The advantage is that the models are created relatively automatically using existing software. The models provide representations that approximately preserve the joint relationships of variables and are easy to apply. The models allow imputation for missing data in contingency tables and for the creation of discrete, synthetic microdata satisfying analytic constraints.

Introduction

Graphical representation of Bayes Nets and other probabilistic relationships date to Lauritzen and Spiegelhalter (1988). They are used extensively in machine learning. For instance, Figure 2 in Getoor et al. (2001) (reprinted below) demonstrates an efficient representation of Census data. 951 parameters are able to represent a potentially large number of cells in a contingency table (7 billion).

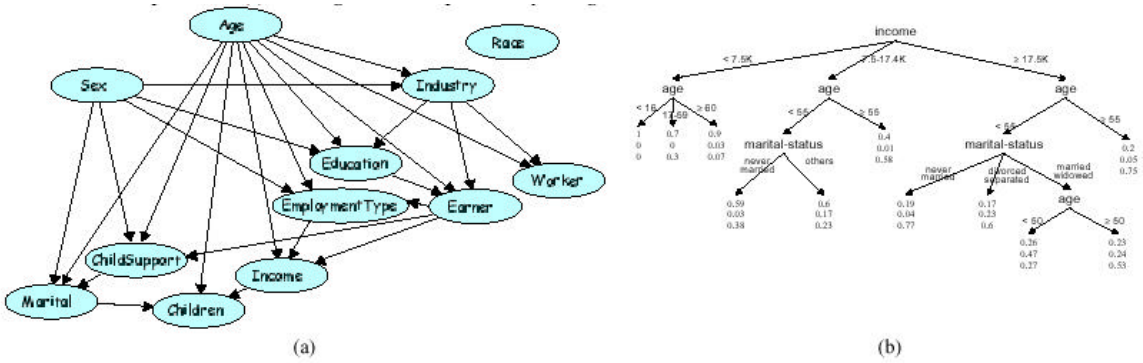

Figure 2: (a) A Bayesian network for the census domain. (b) A tree-structured CPD for the *Children* node given its parents *Income*, *Age* and *Marital-Status*.

Bayes Net software will quickly determine dependency relationships such as those of Figure 2 in Getoor et al. (2001). A mathematical representation is

$$P_B(A_1, \ldots, A_n) = \Pi_{i \le n} \; P_B(A_i \mid \text{Parents}(A_i)). \qquad (1)$$

If as shown in Figure 2, a given variable depends on only a few other variables (i.e., parents), then representation (1) is very efficient. If there is no missing data, then computation of the probabilities in (1) is exceedingly rapid (see e.g., Friedman 1997).

An Alternative to Hot-deck Imputation

Given the representation of Figure 2 (alternatively from equation (1)), one can impute as follows:

B_IMPUTE:

1. Start with any variable A_i having a missing value. If the parents of A_i have nonmissing values, impute the missing value of A_i according to the conditional probabilities of Figure 2.
2. If a parent has a missing value, proceed to its parents. Proceed through parents until reaching the highest point in the Figure. If at the highest point, impute the missing value according to the conditional probabilities

determined by the parents. If there are no parents, impute the probability according to the observed frequency of the variable.

3. Proceed back down the figure filling in all values until the original variables A_i is filled in.
4. Proceed through the remaining variables A_j having missing data as in steps 1-3.

Notes:

1. As with a hot-deck imputation, the above imputation does not deal with non-ignorable nonresponse.
2. Unlike hot-deck imputation, the above imputation procedure preserves joint probabilities.
3. Unlike hot-deck, the above imputation procedure does not rely on matching against representative donors in the set of records.

Imputation that Preserves Edit Relationships

If the contingency table represents only those records that have no missing values and that satisfy all edits, then no edit-failing records will be represented in a figure comparable to Figure 2. All imputations from such a figure will necessarily satisfy all edits.

Note:

1. Because of the parsimony of the representation, the BN (Bayes Net) representation will generally only be a within-epsilon representation. In some situations, it may be possible to impute according the B_IMPUTE procedure in such a manner that the imputation does not satisfy edits. Because the within-epsilon representation covers the overwhelming majority of situations from the original contingency table, most imputations should satisfy edits.

Testing for Nonignorable Nonresponse

It is intuitive that imputation when there are edit restraints will always be nonignorable. Let X_1 and X_2 be two variables that are associated by an edit. That is, the values assumed by variable X_1 are restricted by the values assumed by variable X_2. For instance, if X_1 represents marital status and X_2 represents age, then an edit might be E = {X_1 = married, X_2 <=15}. If a record R has values for variables X_1 and X_2 that coincide with E, then the record fails E. If X_1 has a blank value (either because it is missing originally or because it is blanked after passing through an edit program), then the value that must be imputed depends on the edit E and the value (<= 15) in the age variable X_2. The missingness in variable X_1 is dependent on the value (potentially married) that might be in the variable X_2.

Synthetic Micro-data Satisfying Analytic Constraints

If a large database is given by a contingency table, then it can be approximated by a Bayes Net. If the synthetic data are generated according to the probabilities in the BN, then the synthetic micro-data will preserve most of the analytic properties of the original microdata.

Alternative Representation of a Contingency Table for a Census

Grim et al. (2001) provide a method of representing a contingency table for a census. They represent it as a mixture of multinomial models in which each individual model satisfies a conditional independence assumption. The conditional independence assumption is that the fields are independent given the indicator of being in an individual model. The methods are computationally tractable and provide an alternative to the BN described above.

Software

General WinMine software for Bayesian Networks and, more generally, Dependency Networks is currently available without charge from Microsoft Research (2001).

General latent-class modeling software is described in Winkler (1993). The theory is developed in Winkler (1990, 1993) and an application is described in Winkler (1994). Another application to the problem of text classification with labeled and unlabeled data is described in Winkler (2000).

Additional Remarks

Any parsimonious probabilistic representation in databases having a large number of records (e.g., Davies and Moore 1999, DuMouchel et al. 2000) of data can be used for generating either synthetic data or data that satisfies analytic constraints.

1/ This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

References

Davies, S. and Moore, A. (1999), "Bayesian Networks for Lossless Dataset Compression," *Association of Computing Machinery, Conference of Knowledge Discovery in Data.*

Domingo-Ferrer, J. (ed.) (2002), *Inference Control in Statistical Databases*, Lecture Notes on Artificial Intelligence, Springer: New York.

DuMouchel, W., Volinsky, C., Johnson, T., Cortes, C. and Pregibon, D. (2000), "Squashing Flat Files Flatter," *Association of Computing Machinery, Proceedings of Knowledge Discovery in Data*, 6-15.

Friedman, N. (1997), "Learning Belief Networks in the Presence of Missing Values and Hidden Variables," *in D. Fisher, ed., `Proceedings of the Fourteenth International Conference on Machine Learning', Morgan Kaufmann, San Francisco, CA, pp. 125-- 133*.

Friedman, N. (1999), "The Bayesian Structural EM Algorithm," *in G. F. Cooper & S. Moral, eds., `Proc. Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI '98)', Morgan Kaufmann, San Francisco, CA.*

Getoor, L., Taskar, B., and Koller, D. (2001), "Selectivity Estimation using Probabilistic Models," *Association of* Computing Machinery, Proceedings *of SIGMOD '01* (available at http://robotics.stanford.edu//~getoor/papers/sigmod01.ps).

Grim, J., Bocek, P., and Pudil, P. (2001), "Safe Dissemination of census Results by Means of Interactive Probabilistic Models," *Proceedings of 2001 NTTS and ETK*, Eurostat: Luxembourg, 849-856.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer: New York.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988), "Local Computations with Probabilities on Graphical Structures and Their application to Expert Systems," *JRSS, B 50(2)*, 157-224.

Little, R. J. and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, John Wiley: New York.

Microsoft Research (2001), "WinMine Tookit," http://research.microsoft.com/~dmax/WinMine/tooldoc.htm.

Winkler, W. E. (1990), "On Dykstra's Iterative Fitting Procedure," *Annals of Probability*, **18**, 1410-1415.

Winkler, W. E. (1993), "Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 274-279 (report 93/12 available at http://www.census.gov/srd/www/byyear.html).

Winkler, W. E. (1994), "Advanced Methods for Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 467-472 (longer version report 94/05 available at http://www.census.gov/srd/www/byyear.html).

Winkler, W. E. (2000), "Machine Learning, Information Retrieval, and Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 20-29. (also available at http://www.niss.org/affiliates/dqworkshop/papers/winkler.pdf).