Working Paper No. 5
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)**

Joint ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (i): New theories and emerging methods

BAYESIAN NETWORKS REPRESENTATIONS OF CONTINGENCY TABLES FOR REDUCING DISCLOSURE AND PRESERVING THE ACCURACY OF SUFFICIENT STATISTICS

Invited Paper

Submitted by the U.S. Bureau of the Census, United States[1]

[1] Prepared by William E. Winkler (william.e.winkler@census.gov) and Yves Thibaudeau (yves.thibaudeau@census.gov).

Bayesian Networks Representations of Contingency Tables for Reducing Disclosure and Preserving the Accuracy of Sufficient Statistics

Yves Thibaudeau and William E. Winkler

Abstract

The paper explores the use of Bayesian networks to simulate discrete data reported in a contingency table in a manner that reconstructs the contingency table with reduced risk of disclosure, relative to the original table. The object of the investigation is the degree of integrity of the network in maintaining the accuracy of certain sufficient statistics through the simulation process. While with general log-linear models it is often possible to guarantee a realistic simulation of the interactions between specific variables, the parsimony of Bayesian networks implies a reduced level of control in reproducing interactions. We experiment with techniques to construct Bayesian networks and compute the associated conditional probabilities in order to simulate entire contingency tables, or selected cells in the tables. At the same time we measure the level of accuracy of certain sufficient statistics computed from the modified table, relative to the original table. We attempt to delineate a strategy for identifying Bayesian networks for simulating contingency tables that maintain a relatively good accuracy for certain sufficient statistics, while also reducing the risk of disclosure inherent to the original table

1. Introduction

The objective of the paper is to propose an alternative computation method for the transition probabilities of a Bayesian network, as it is defined based on the running intersection property. The general context of method is that of current effort to prevent information disclosure associated with contingency tables while also maintaining a level of information that is calibrated with an engine based on a reduced dimension sufficient statistic in connection with a log-linear model.

Our medium to build method that limit disclosure, either directly or through Bayesian networks, is the definition of a family of likelihoods and partial likelihoods, which serves to represent to the underlying model at the heart of the engine. The particular representation of the likelihoods must be such that it allows the identification of reduced dimension sufficient statistics that will be preserved through the disclosure limitation process.

2. Homogeneous Association Models for Bayesian Network

Bayesian networks can be represented by acyclic directed graphs flowing trough "cliques" of nodes (Lauritzen, Spiegelhalter 1988). The "running intersection" property of a triangulated graph along with the direction of graph determines the inheritance process through the nodes. We associate each node in the network with one dimension of a contingency table. The simplest non-trivial situation is a single triangle of nodes, each representing the occurrence or non-occurrence of an event. This node structure can be

represented by a two-by-two-by-two contingency table that we model with a homogeneous-association (Schafer 1997) log-linear model. In this representation, one dimension of the three-dimensional contingency tables represents the child of the two other nodes, which are then its "parents".

To represent the likelihood of the homogeneous associations log-linear model corresponding to such a contingency table, let $\left\{ R_{i,j,k} \right\}$ be the set counts for the cells of the contingency table. Let $y_{1,1}, y_{1,2}, y_{2,1}, 1 - y_{1,1} - y_{1,2} - y_{2,1}$ be the marginal probabilities corresponding to the two first dimensions, and let $g_{1,1}, g_{1,2}, g_{2,1}$ be the conditional probabilities of the events corresponding to the third dimension in the conditional space generated by the first three cells defined by the first two dimensions respectively (the probability conditional on the fourth cell is implicitly defined as a function of $g_{1,1}, g_{1,2}, g_{2,1}$. The joint likelihood for these parameters is:

$$
\begin{aligned}
&L\left(y_{1,1}, y_{1,2}, y_{2,1}, g_{1,1}, g_{1,2}, g_{2,1}\right) = \\
&= y_{1,1}^{R_{1,1,1}+R_{1,1,2}} \\
&\times y_{1,2}^{R_{1,2,1}+R_{1,2,2}} \\
&\times y_{2,1}^{R_{2,1,1}+R_{2,1,2}} \\
&\times \left(1-y_{1,1}-y_{1,2}-y_{2,1}\right)^{R_{2,2,1}+R_{2,2,2}} \\
&\times \left(g_{1,1}\right)^{R_{1,1,1}+R_{2,2,2}} \left(1-g_{1,1}\right)^{R_{1,1,2}+R_{2,2,1}} \\
&\times \left(g_{1,2}\right)^{R_{1,2,1}+R_{2,2,1}} \left(1-g_{1,2}\right)^{R_{1,2,2}+R_{2,2,2}} \\
&\times \left(g_{2,1}\right)^{R_{2,1,1}+R_{2,2,1}} \left(1-g_{2,1}\right)^{R_{2,1,2}+R_{2,2,2}} \\
&\times \left[\begin{array}{l}\left(1-g_{1,1}\right)g_{1,2}g_{2,1} \\ \qquad + g_{1,1}\left(1-g_{1,2}\right)\left(1-g_{2,1}\right)\end{array}\right]^{-\left(R_{2,2,1}+R_{2,2,2}\right)}
\end{aligned}
\qquad (1)
$$

Based on (1) a natural sufficient statistic for the joint parameters is

$$
T = \left(\begin{array}{l} R_{1,1,1}+R_{1,1,2}, R_{1,2,1}+R_{1,2,2}, R_{2,1,1}+R_{2,1,2}, R_{2,2,1}+R_{2,2,2}, \\ R_{1,1,1}+R_{2,2,2}, R_{1,2,1}+R_{2,2,1}, R_{2,1,1}+R_{2,2,1} \end{array}\right)
\qquad (2)
$$

This statistic is minimal sufficient in the sense that its dimensionality is equal to the number of degrees of freedom of the model (number of free parameters) plus one additional dimension to account for the total (intensity of the process). To define the transition probabilities of the Bayesian network, we use the MLEs of the conditional probabilities $g_{1,1}, g_{1,2}, g_{2,1}$. Because of the reduced dimensionality of T relative to that of the table, we are guaranteed some disclosure protection.

3. Using the MLE and the Sufficient Statistic to Build Substitute Table

An more traditional alternative to using the sufficient statistic to construct a Bayesian network, is to use it to construct a value-preserving, in terms of the sufficient statistic, substitute table that exhibits different values for some cells, in particular for sensitive cells. The first step is to enumerate all legitimate values of the vector of response variables $R = \left\{ R_{i,j,k} \right\}$ such that $T = t$, which can be done by solving a system of linear equations. Then the second step is to construct the following conditional probability function:

$$P\left(R = r \middle| T = t \right) = \frac{P\left(R = r \right)}{\sum_{\left\{ s \middle| t(s) = t \right\}} P\left(R = s \right)}; \qquad t\left(r \right) = t \tag{3}$$

In (3) the variables are all vectors of the appropriate dimensions. Then the conditional probability distribution in (3) can be simulated with the parameter values replaced by their MLE and the outcomes that provide minimal disclosure are retained. This approach guarantee a full recovery of the original values of T, and thus of the MLE, from any substitute table produced this way.

4. Other Sufficient Statistics

Sufficient statistics such as that defined in (2) can be obtained for three-dimensional contingency tables of any size. In the context of Bayesian networks one can increase the number of parents involved in the causality process. That is, increase the dimensionality of the underlying contingency table. Thibaudeau (2000) suggests a way to construct such sufficient statistics for contingency tables of dimension higher than 3.

5. Conclusion

We have explored a method based on parametric statistics to identify a sufficient statistic that summarize the information contained in a higher dimensional contingency table given that some log-linear model underlies the process that produced the table. The sufficient statistics is of reduced dimension relative to the degree of freedoms of the table itself. Such a sufficient statistic can be used naturally as a substitute for a transition probability in a Bayesian network to limit the amount of information available through the transition, based on the running intersection property. The sufficient statistic can also serve as conditional value in the construction of a table that maintains this value for the sufficient statistic, but has different cell values in order to prevent disclosure. Simulation work, as well as the exploration of higher-dimensional cases remains to be done.