Causal Discovery from Population-Based Infant Birth and Death Records

Subramani Mani and **Gregory F. Cooper**
Center for Biomedical Informatics, Suite 8084, Forbes Tower, 200 Lothrop Street,
University of Pittsburgh, Pittsburgh PA 15213, Ph: 412-647-7113; Fax: 412-647-7190
{mani,gfc}@cbmi.upmc.edu

Introduction

The most useful explanation of a phenomenon is often a description of the underlying causal processes (Salmon 1998). This is particularly true in the domain of medicine where identification of the causal factors of a disease influence treatment planning and development of intervention strategies for disease prevention and control. The study described here focuses on causal discovery from observational data related to infant mortality in the United States.

Methods

Infant Birth and Death Dataset

We used the U.S. Linked Birth/Infant Death dataset for 1991. This consists of information on all the live births in the United States for the year 1991. It also has linked data for infants who died within one year of birth. We selected a random subset of 41,155 cases (1% of the total sample) for use in the current study.

An Algorithm for Causal Discovery

In this section, we introduce a causal discovery algorithm called LCD2. LCD2 assumes the following:

Assumption 1: Causal relationships are represented using Bayesian Networks.

Assumption 2: Variables W and Y test as being independent given variable X, if and only if W and Y are d-separated (Pearl 1991) given X.

Assumption 3: There is a variable W that is not caused by any other measured variable in the dataset.

Given Assumptions 1–3 above and using a set of *independence* and *dependence* tests detailed in (Cooper 1997), the LCD2 algorithm explores a database for possible causal relationships of the following form $W \rightarrow X \rightarrow Y$. This is a causal model in which W causes X, and X causes Y. If the data suggests the presence of such causal relationships, LCD2 outputs that X causally influences Y, and it displays the probability distribution of Y given X. The time complexity of LCD2 is $O(mn^2)$, where m is the number of records in the database and n is the number of variables. We implemented LCD2 in the PERL programming language. For W we used *Race of the mother.*

Results

When applied to the infant birth and death dataset, LCD2 output nine purported causal relationships. Table 1 shows the probability distributions associated with one of the relationships: *Birth-weight* $\rightarrow$ *Infant Outcome.*

Table 1: Conditional Probability Table of Infant Outcome At One Year given Infant Birth Weight

Birth Weight	Infant outcome at one year	
	Survived	Died
<1500 gms.	0.713*	0.287
1500–2499 gms.	0.977	0.023
$\geq$ 2500 gms.	0.997	0.003

*The probability that Infant outcome at one year equals Survived *given* that Infant Birth Weight *is* <1500 grams.

Discussion and Conclusion

Out of the nine causes discovered in the infant birth and death dataset, eight appear plausible. Due to space limitations we showed only one of them here. From Table 1 we can see that as the birth weight increases from less than 1500 grams to 1500–2499 grams and then to 2500 or more grams, the probability of survival increases from 0.713 to 0.977 and further to 0.997.

In summary, for this dataset, the LCD2 algorithm appears to be outputting relationships that on the whole are plausibly causal.

References

Cooper, G. F. 1997. A Simple Constraint-Based Algorithm for Efficiently Mining Observational Databases for Causal Relationships. *Data Mining and Knowledge Discovery* 1:203–224.

Pearl, J. 1991. *Probabilistic Reasoning in Intelligent Systems.* San Francisco, California: Morgan Kaufmann.

Salmon, W. C. 1998. *Causality and Explanation.* New York: Oxford University Press.