On Retaining Intermediate Probabilistic Models When Building Bayesian Networks

Prashant J. Doshi and **Lloyd G. Greenwald**
Department of Mathematics and Computer Science
Drexel University

John R. Clarke
Department of Surgery
MCP-Hahnemann University

Introduction

The process of building a Bayesian network may occur in stages, in which intermediate Bayesian networks are built during preliminary processing and then used in the construction of further Bayesian networks. For example, in (Doshi, Greenwald, & Clarke 2001) we describe a way to use Bayesian networks to model and correct errors in noisy datasets. The corrected datasets are then used in (Doshi 2001) to build predictive Bayesian networks. Through this process we built networks that capture probabilistic relationships between 412 fields of data from 169,512 patients admitted to trauma centers in Pennsylvania and registered in the *Pennsylvania Trauma Systems Foundation Trauma Registry* between 1986 and 1999.

In the process mentioned above, intermediate Bayesian networks were used to find the *most likely* values for fields found to have errors. These most likely values were then used in the cleansed dataset. However, in the subsequent process of building Bayesian networks from this dataset, we questioned whether or not these intermediate networks used in error correction should have been retained. In other words, we wanted to understand the tradeoffs involved in retaining the distributional information summarized in each error-correction network rather than just retaining the most likely value for each corrected field. This question can be generalized to any process of building a Bayesian network in stages. This note describes preliminary work to understand these issues.

An important component of this staged network building process is that common variables are represented from one stage to the next. In data cleansing, variables used to query for error distributions are the same variables that are used as evidence variables in the final predictive network. Furthermore, the context variables used to model errors are also represented directly in the final network. Retaining distribution information can be accomplished by employing networks from early stages within the subsequent networks. Common variables limit the potential blow-up in network size.

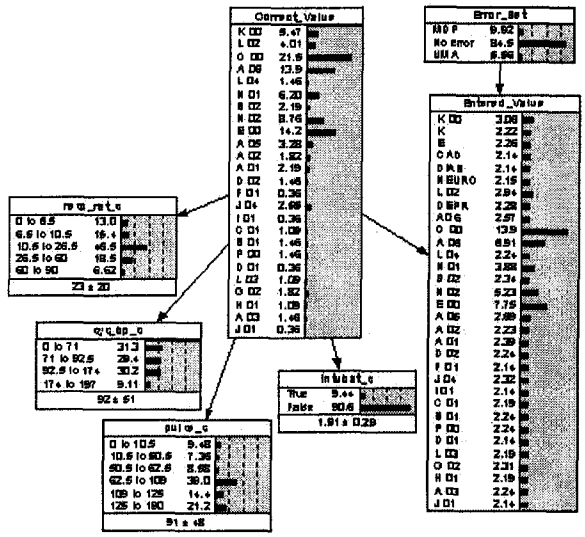

Figure 1: Error correction Bayesian network

Potential benefits of retaining distributional information include improved predictive accuracy of the resulting network. Additional benefits might be found in efficiency of learning parameters of subsequent networks using the intermediate networks as a form of prior knowledge. Disadvantages include potential blow-up in network size and the subsequent effects on parameter learning and inference.

In the following section we give preliminary results on improved predictive accuracy due to retaining intermediate models. Our experimental setup consists of a Bayesian error correction model and a patient outcome prediction network, both built directly from real patient data and expert knowledge of the domain and database.

Experimental Analysis

In this section we present experimental setups consisting of a Bayesian error correction model and a patient outcome prediction network. The structure of these networks is elicited from a medical expert and probabilistic information is learned from data extracted from a subset

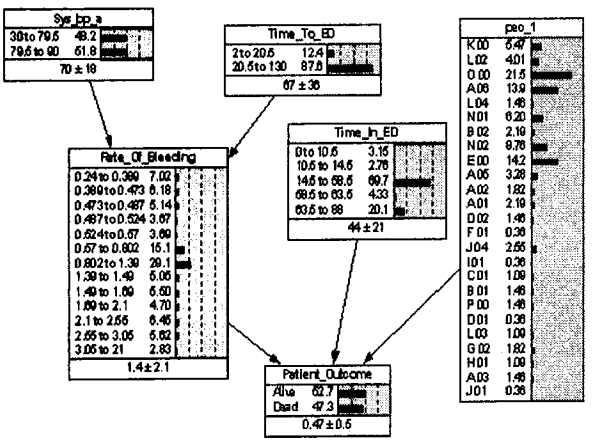

Figure 2: Patient outcome prediction Bayesian network

of our trauma care database.

The data used to train and test the networks is obtained from a study we conducted to show that the time to laparotomy for intra-abdominal bleeding from trauma does affect survival for delays upto 90 minutes (Clarke *et al.* 2001). The data represents patients who were admitted to the trauma wards in 27 institutions across the state of Pennsylvania for intra-abdominal bleeding and who were moved to the operating room for laparotomy within 90 minutes. 496 such patients were identified from the data and we looked at 12 variables for each patient. The isolated data was randomly divided into training and test sets, each containing an equal number of records. The conditional probability tables for the networks were learned using the training set and inference was carried out using the test sets.

The error correction model shown in Figure 1 employs a context-driven error correction approach developed in (Doshi, Greenwald, & Clarke 2001). The model displays a probability distribution over the Correct_Value for the data field *patient preexisting conditions* given the errant Entered Values and optionally, the type of error listed in Error Set and the context variables: respiratory rate, systolic blood pressure, pulse and intubation at scene of injury. A likelihood measure may then be computed from this probability distribution and entered as findings for a node in the patient outcome prediction network (Figure 2) or the distribution may be directly used in the prediction network.

Figure 2 is a patient outcome prediction network that relates patient preexisting conditions(pec_1), systolic blood pressure on arrival in the trauma ward, time to ED (Emergency Department), time in ED, and rate of bleeding to Patient_Outcome. The network may be used to infer probability of survival (Patient_Outcome) from a given set of evidence values.

Prediction using the Most Likely Value The most likely value for the variable Correct_Value for each test instance was inferred from the network in Figure 1. The network in Figure 2 was then tested for Patient_Outcome using test cases that contained the most likely values as findings for the node pec_1 along with test data for the other observed variables. The resultant confusion matrix is given below.

Predicted		Actual
Alive	Dead	State
139	17	Alive
41	50	Dead

Error rate = 23.48%

Prediction using the Probability Distribution In order to directly make use of the probability distributions inferred from the error model in the patient prediction model, we combined the two networks by merging the nodes pec_1 and Correct_Value and copying all other nodes. Findings from the test instances were entered only for the observed nodes i.e. no findings were entered for the variable Correct_Value, and the network was tested on Patient_Outcome. We observed a reduction of 2.02% in the error rate compared to the error rate in the previous section as is evident from the confusion matrix shown below.

Predicted		Actual
Alive	Dead	State
132	24	Alive
29	62	Dead

Error rate = 21.46%

Discussion These results demonstrate improved predictive accuracy due to retaining intermediate probabilistic models. We are analyzing representative Bayesian inference algorithms to provide a careful theoretical understanding of these initial results. We are also currently quantifying the tradeoffs in inference time. Furthermore, we are studying the effects of removing the data cleansing stage entirely and comparing predictive networks built directly from noisy data with those built by using expert knowledge to build intermediate error correction models.

References

Clarke, J. R.; Trooskin, S. Z.; Doshi, P. J.; Greenwald, L. G.; and Mode, C. J. 2001. Time to laparotomy for intra-abdominal bleeding from trauma does affect survival for delays up to 90 minutes. *Journal of the American Association of Trauma Surgeons.*

Doshi, P. J.; Greenwald, L. G.; and Clarke, J. R. 2001. Context-driven methods for detecting and correcting errors in large databases. In *Under Review.*

Doshi, P. J. 2001. Effective methods for building probabilistic models from large noisy data sets. Masters Thesis, Department of Mathematics and Computer Science, Drexel University.