Analysis of a causal modeling approach: a case study with an educational intervention

Dovan Rai and Joseph E. Beck
{dovan, josephbeck }@wpi.edu
Computer Science Department, Worcester Polytechnic Institute

1 Introduction

This paper explores the application of causal models to understanding data generated from a computer tutor. Mily's World (http://users.wpi.edu/~dovan/coordinates.html) is a flash-based learning environment for coordinate geometry and featuring *game-like properties* such as a cover story and pictures. We were primarily interested in what class of students benefitted from this type of intervention, as well as which students preferred this style of instruction to traditional materials. Fifty eight students used the tutor and we collected survey data and their log records from the tutor. We analyzed the data using Tetrad (http://www.phil.cmu.edu/projects/tetrad), free software designed for causal modeling.

2 Causal model search

We used the PC algorithm in Tetrad, which is designed to search for causal explanations of observational or mixed observational and experimental data. The causal model thus obtained is shown in Figure 1 where the rectangular nodes are the data of each student.

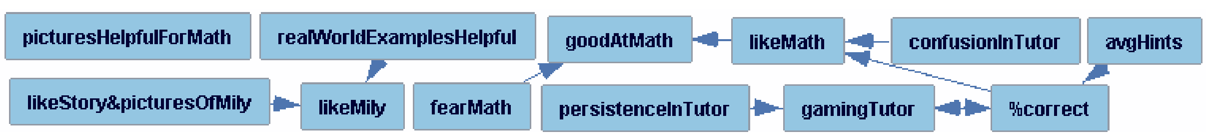

Figure 1 Search Model: causal model after making PC search

Based on our domain knowledge, we hand-crafted a causal model shown in Figure 2, where we added four latent nodes (oval nodes) that we believe are influencing the observables.

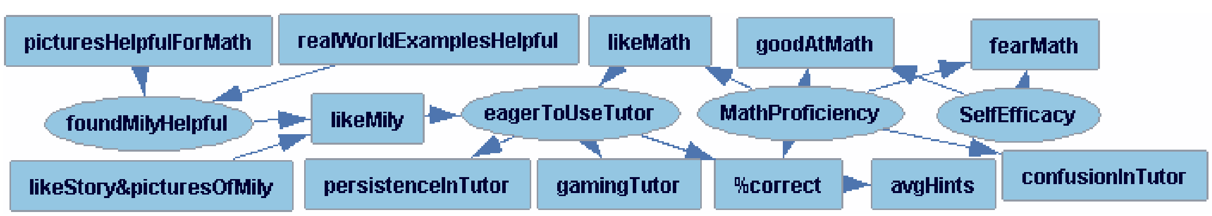

Figure 2 Hand-crafted model (latent variables are ovals)

We also generated a correlation graph where a link between two nodes indicates they are reliably correlated with $P<0.05$. We then compared the PC causal model and the correlation graph with our hand-crafted model. For each link in the hand-crafted model, if the automatically generated model had it it was a true positive; if the link was missing it was a false negative. Similarly, if the model has correctly identified the absence of link, that would be a true negative. The causal model is more stringent than correlation graph as it would put a link between nodes only if they retain their association after controlling for all other nodes. We found that the correlation graph has more false positives whereas the causal model is more susceptible to false negatives.

2.1 *True positive with correct direction and true negatives*

By exploiting conditional dependencies, the PC model correctly identifies true positives with correct direction (LikeStory&picturesOfMily→ likeMily ← realWorld ExamplesHelpful) and true negatives (link likeMath–avgHints is gone once controlled for "% correct"). This ability to automatically partial out other influences is difficult, at best, to replicate in traditional statistics packages.

2.2 *False negatives: weaker statistical power due to small sample size*

When we have small sample size, doing partial correlations can give false negatives due to limited statistical power. Having more samples reduces false negatives without adding false positives. Multicollinearity is an extreme case, where we might falsely conclude that there is no linear relationship between an independent and a dependent variable. For example: picturesHelpfulForMath is correlated with both likeMily and LikeStory&picturesOfMily. But since, likeMily and LikeStory&picturesOfMily are highly correlated between themselves (.471**), picturesHelpfulForMath is conditionally independent to both of them (see Figure 1).

2.3 *Search with domain knowledge*

To overcome the problem of multiple "Markov equivalent" graphs that can be built from the same data, we add domain knowledge to direct our search to pick the most compatible model.

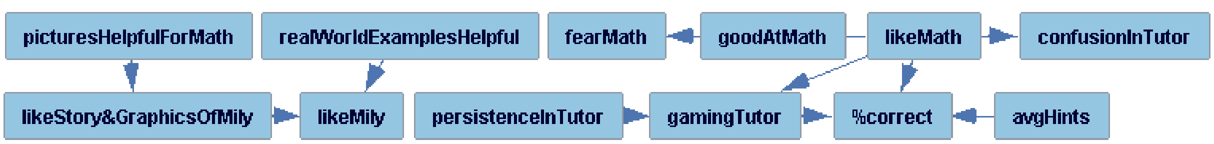

Figure 3 Causal model with domain knowledge

We see from Figure 2 and Figure 3 that adding domain knowledge not only fixes the arrow orientations (likeMath →%correct), but also adds new edges such as likeMath→gamingTutor. One interesting finding is that adding domain knowledge has fixed the problem of multicollinearity (picturesHelpfulForMath→LikeStory&picturesOfMily) as adding temporal knowledge restricts nodes to only influence things which occurred later.

3 Conclusions

In this paper, we have presented a case study of applying causal modeling, using the Tetrad software, to understand what factors influence how students respond to our educational intervention. We found that a problem that arises from having a small sample results in more false negatives in our causal model. That is, there are true relationships that we lack the statistical power to detect. We also found that by adding domain knowledge, we are not only able to correct the arrow orientations but we can also overcome issues such as multicollinearity to come up with the most plausible model from the set of equivalent models.