Evaluating Causal Bayesian Networks for Diabetes Risk Assessment: A Comparative Analysis of Parameterization and Inference Techniques

Ehsan Ramezani
Master's Degree in Artificial Intelligence, University of Bologna
{ ehsan.ramezani}@studio.unibo.it

April 17, 2024

Abstract

This mini-project aims to construct a foundational model structure of a Bayesian network to address variables influencing diabetes.Using various parameterization methods, I demonstrated that there is a tiny difference between their conditional probability distribution (CPDs). Additionally, both exact and approximate inference techniques are employed to evaluate the practical application of Bayesian networks in resolving inferential queries about diabetes.

I found that there are varied correlations between variables that have effects on diabetes outcomes. My comparative analysis of different inference methods focused on assessing runtime efficiency and probability scores, revealing key insights into the predictive capabilities of Bayesian networks for healthcare applications.

Introduction

Domain

Diabetes is a major global health issue that impacts millions of people and places a significant burden on healthcare systems. The chosen study, ”Investigating the validity of structure learning algorithms in identifying risk factors for intervention in patients with diabetes,”(Zahoor et al. 2024)utilizes a variety of structure learning algorithms to model Causal Bayesian Networks (CBNs) from the BRFSS-2015 dataset, with the goal of comprehending the causal pathways that influence the progression of diabetes.

The paper presents three Directed Acyclic Graphs (DAGs)—High Confidence, Moderate Confidence, and Low Confidence—that illustrate the relationships between 22 variables associated with diabetes. These variables are categorized into non-modifiable risk factors (such as age and sex), modifiable risk factors (such as BMI and blood pressure), and medical conditions (such as diabetes status). I chose the Low Confidence DAG because it has edges between variables with low confidence, which provides a comprehensive overview of potential connections, and it allows individuals without domain knowledge to ask and evaluate different queries about variables with low confidence as well as those with high confidence, allowing us to observe the effect of these edges on determining whether they have an impact on diabetes.

Aim

As stated in the study ”Investigating the validity of structure learning algorithms in identifying risk factors for intervention in patients with diabetes,” the goal of this project is to use the Low Confidence Bayesian network and conduct experiments using the concepts taught in the course. Additionally, do an experiment applying the Maximum Likelihood Estimator and Bayesian Estimator to parameterize the network model. The CPDs generated by these estimators on dataset are then evaluated. Furthermore, posing relevant questions based on the specified structure mentioned in the academic article and employing different queries in order to achieve important findings. finally, another objective of this study is to do a comparative analysis of the runtime and error associated with approximate inference versus the exact inference approach.

Method

I used pgmpy[1] library features to set up our network and perform queries. To determine if variations in estimate techniques, queries, parameters, or sample size caused differences in accuracy, we altered evidence nodes and inference methods. Since there were no CPDs in the publication, I used the pandas[2] package to interact with the dataset and learn parameters from it.

Results

- The importance of having adequate data in sensitive healthcare scenarios, such as diagnosing diabetes, underscores the significance of dataset size.
- Bayesian estimators, in comparison to maximum likelihood estimator, demonstrate greater robustness in cases where datasets are limited, but since there is sufficient data, their precision is close.
- Understanding the information flow between variables in large networks (conditional independence), whether there's evidence or not, along with identifying their Markov blankets, provides valuable insights into complex systems.

[1] https://pgmpy.org/

[2] https://pandas.pydata.org/

- The run time of the Pgmpy library's sample generation function is determined by factors such as network topology, the number of variables, and the edges connecting them, as demonstrated by the time required to generate 10,000 samples using the rejecting sample method across various scenarios. Furthermore, making use of `get_independencies()` and `get_assertions()` in this network consumes a significant amount of time and CPU resources, surpassing the capabilities of my personal laptop or the free version of Google Colab.

Model

Here's a table with the variables used to build the networks and provided in the dataset:

Variable	**States**
Diabetes.binary	{0 No, 1 Yes}
HighBP	{0 No, 1 Yes}
HighChol	{0 No, 1 Yes}
BMI	{0 0-24, 1 25-39, 2 $\geq$ 40}
HeartDiseaseOrAttack	{0 No, 1 Yes}
CholCheck	{0 No, 1 Yes}
Stroke	{0 No, 1 Yes}
Smoker	{0 No, 1 Yes}
Fruits	{0 No, 1 Yes}
Veggies	{0 No, 1 Yes}
HvyAlcoholConsump	{0 No, 1 Yes}
AnyHealthcare	{0 No, 1 Yes}
NoDocbcCost	{0 No, 1 Yes}
MentHlth	{0, 1, 2}
PhysHlth	{0, 1, 2}
DiffWalk	{0 No, 1 Yes}
Sex	{0 Female, 1 Male}
Age	{1, 2, 3, 4, 5, 6}
Income	{1, 2, 3, 4}
Education	{1, 2, 3}
GenHlth	{1, 2, 3}
PhysActivity	{0 No, 1 Yes}

Table 1: Description of variables and their states

The juptyer notebook file contains more relevant information on variables besides the network's structure and the process of building it.

Analysis

Experimental setup and Results

Comparison of Estimation Methods

here are tables showing differences in CPDS between two estimators:

Education

State	MLE	Bayesian
Education(1)	0.616623	0.616658
Education(2)	0.284721	0.284723
Education(3)	0.698656	0.698641

Smoker

State	MLE	Bayesian
Smoker(0)	0.556831	0.556829
Smoker(1)	0.443169	0.443171

here are tables showing differences in CPDs between variable elimination (exact inference) and rejecting samples (approximate inference) :

$$P(\textbf{Diabetes\_binary} \mid \textbf{Income} = 4, \textbf{Education} = 3)$$

Diabetes_binary	**Exact**	**Approximate**
Negative (0)	0.8910	0.905
Positive (1)	0.1090	0.095

$$P(\textbf{Diabetes\_binary} \mid \textbf{Income} = 2, \textbf{Education} = 2)$$

Diabetes_binary	**Exact**	**Approximate**
Negative (0)	0.7753	0.6645
Positive (1)	0.2247	0.3355

$$P(\textbf{Diabetes\_binary} \mid \textbf{Income} = 1, \textbf{Education} = 1)$$

Diabetes_binary	**Exact**	**Approximate**
Negative (0)	0.5984	0.6009
Positive (1)	0.4016	0.3991

Conclusion

In my study for this project, I discovered that variables that look insignificant can be impactful on diabetes, and vice versa; alcohol intake even benefits in some circumstances, contrary to my expectations. Furthermore, having an extensive network and enormous data gathering improves the accuracy of work in sensitive areas such as medical care and health. However, the complexity of the network and the size of the dataset are associated with a trade-off between accuracy and available resources; where I was limited by a lack of suitable computing resources, removing this limitation will allow for better variable analysis.for example, studying low-to-medium correlation between variables and their occurance in independence assertions.

Links to external resources

Link to Dataset: Diabetescategorised.csv

References

[Zahoor et al. 2024] Zahoor, S.; Constantinou, A. C.; Curtis, T. M.; and Hasanuzzaman, M. 2024. Investigating the validity of structure learning algorithms in identifying risk factors for intervention in patients with diabetes. *arXiv preprint arXiv:2403.14327*.