ByeBayes, a Bayesian Network for Death Probability

Andrea Zecca, Giuseppe Carrino, Stefano Colamonaco
Master's Degree in Artificial Intelligence, University of Bologna
{ andrea.zecca3, giuseppe.carrino2, stefano.colamonaco}@studio.unibo.it

October 15, 2022

Abstract

This mini-project aims to create a medium-size Bayesian network to forecast death probabilities of individuals affected by serious injuries or several infections. Behind this project there are several scientific studies focused on the main death causes. We used two sundry models composed of about 14/15 nodes, each of them modelling a different aspect of the health of an individual, with the only purpose to forecast its probability of death.

Introduction

Domain

ByeBayes is a project which implements a few models for death causes probabilities, given a set of age ranges and subject gender. The project has been constructed following research papers about death causes in order to get realistic probabilities, so percentages formulation didn't require huge effort (Roth et al. 2018). Otherwise, the challenge has been the data interpretation and the model drawing, given numerical information, in order to develop an interesting research job.

Aim

The purpose of *ByeBayes* is to implement a original Bayesian Network given only numerical data, manually constructing different kind of models and comparing differences between them, investigating how different implementative choices influence inference and variables dependencies.
Furthermore, it was essential to study constructed models with concepts studied during classes, such as Markov Blanket and (targets, evidences)-type queries, in order to deepen knowledge about them.
Also, the aim is to put an accent on network construction, developing a manually-constructed dataset in order to see the precision of sampling algorithm (sampling from an empty network) used to create the dataset.

Method

The whole project has been written with Python, in particular:

- .BIF[1] files were constructed for variables and their own probabilities, using research papers-based percentages. Two .bif files were made: in particular, one also considered COVID-19 variables and its percentages(Poletti et al. 2021);
- A script has been written for manual .CSV dataset construction based on .bif files probabilities;
- Three models have been built - two based on .bif files, one on the dataset - using pgmpy library, considering a node for every variable and holding a CPD table for each one;
- Inference queries, Markov Blankets and dependencies have also been calculated using pgmpy, showing results through tables and graphs drawing.

Results

Thanks to this project, we understood the complexity of modelling a Bayesian Network of medium dimensions, because of the dependencies between variables. Also, we analyzed the effects on inference of modifying a network adding a new variable (*Covid*).

Model

Here are the two main models we rely on within the project, both built manually. The difference between the two models is that in the second an extra *Covid* node (the one in red) was added to analyze the variations that the spread of *SARS-CoV-2* has brought in the area of respiratory tract infections and number of deaths. The models contain the following nodes, divided as follow:

- Personal info, containing the nodes: *Age range* and *Gender*;
- Probability to contract the disease/infection described by the variable itself, containing the nodes: *Respiratory infection*(Ueno et al. 2019), *Malignancy*(White et al. 2014), *Cardiovascular disease*(Miao et al. 2021), *Poisoning*(Center 2020), *Nature force*(Ritchie and Roser 2014), *Fall*(Stevens and Sogolow 2005) and *Covid (only in the Covid-model)*(Poletti et al. 2021);
- Probability to take damage from something in between described in the previous point, containing the

[1] https://www.cs.washington.edu/dm/vfml/appendixes/bif.htm

nodes: *Self-harm violence*(Ritchie, Roser, and Ortiz-Ospina 2015), *Other injuries*, *Sickness*, *Transport incident*, *Other sickness* and *Unintentional injuries*, (Denfeld et al. 2022);

- Probability of dying, containing the node *Death* (Roth et al. 2018).

All of the nodes are discrete random variable with range in [*true*, *false*]. The only differences are represented by *Age range* and *Gender* that have range respectively in [*young, adult, old*] and [*male, female*]. Primarily, in order to build our model, we followed the *causal order* trying to figure out which nodes had directly influence on which others. Secondly, we tried to assign the probability based on the papers we used to carry out our research.

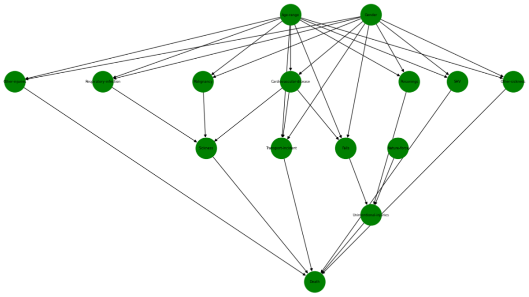

Figure 1: Standard Model

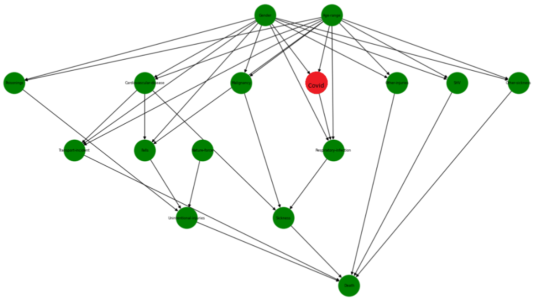

Figure 2: Covid Model

Analysis

Experimental setup

The experiments we carried out wanted to highlight differences between dataset-generated model and .bif one, other than comparing Standard Model and Covid one.
Main Covid queries have been made on both these last two models, in order to understand the impact of adding a new node to the Standard Model. Also, we made queries without Covid as evidence in order to understand the difference between *evidence* and *existence* of a variable in a model.
In order to compare Standard and Dataset Model some random queries have been made. Finally, we analyzed queries about gender and age differences, given injuries or illnesses.

Results

We observed the dependency of the variables, highlighting that Nature-Force was independent from every other node. About inference, we discovered that the dataset-generated model had a very low error (of one magnitude order lower than the actual result) compared to the bif-generated one, demonstrating the correctness of the dataset generation and, also, of the model building.
Covid-related queries have shown, instead, the high impact of the virus on Respiratory-infection, Sickness and Death probability (more than doubling the probability of getting a respiratory infection if infected by Covid-19). Also queries with no evidences about Covid (or even setting Covid as False) highlighted differences on probabilities comparing Standard Model and the one with Covid node added, because of the dependency of variables to the Covid one.

Conclusion

The project can be said to be successfully completed as the objectives have been achieved, the models built can be said to be correct and the comparisons were interesting.

Limitation

The main limit of our project is that it includes information taken from multiple studies on different zones of the world and in different periods of time making a proportion to interpolate the most correct values possible. Of course this means that the information used are not totally reliable.
Another limitation is given by the amount of data analyzed. In fact, the dataset only contains data from 1 million individuals, whilst the data contained in the bif files refer to the entire population.

Future work

Not all the information present on the main document has been used for this project, it would be possible to model all the missing information to obtain an even more precise and complete model.

Links to external resources

- This is a link to the GitHub repository related to the project: ByeBayes;
- This is a link to the Dataset generated by us and uploaded on Kaggle: Dataset.

References

[Center 2020] Center, N. C. P. 2020. Poison statistics. *National Capital Poison Center*. https://www.poison.org/poison-statistics-national.

[Denfeld et al. 2022] Denfeld, Q. E.; Turrise, S.; MacLaughlin, E. J.; Chang, P.-S.; Clair, W. K.; Lewis, E. F.; Forman, D. E.; Goodlin, S. J.; in Older Populations Committee of the Council on Clinical Cardiology, A. H. A. C. D.; on Cardiovascular, C.; on Lifestyle, S. N. C.; Health;, C.; and Council, S. 2022. Preventing and managing falls in adults with cardiovascular disease: a scientific statement from the american heart association. *Circulation: Cardiovascular Quality and Outcomes* 10–1161.

[Miao et al. 2021] Miao, Q.; Zhang, Y.-L.; Miao, Q.-F.; Yang, X.-A.; Zhang, F.; Yu, Y.-G.; and Li, D.-R. 2021. Sudden death from ischemic heart disease while driving: cardiac pathology, clinical characteristics, and countermeasures. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research* 27:e929212–1.

[Poletti et al. 2021] Poletti, P.; Tirani, M.; Cereda, D.; Trentini, F.; Guzzetta, G.; Sabatino, G.; Marziano, V.; Castrofino, A.; Grosso, F.; Del Castillo, G.; Piccarreta, R.; Andreassi, A.; Melegaro, A.; Gramegna, M.; Ajelli, M.; Merler, S.; and Force, A. L. C.-. T. 2021. Association of Age With Likelihood of Developing Symptoms and Critical Disease Among Close Contacts Exposed to Patients With Confirmed SARS-CoV-2 Infection in Italy. *JAMA Network Open* 4(3):e211085–e211085.

[Ritchie and Roser 2014] Ritchie, H., and Roser, M. 2014. Natural disasters. *Our World in Data*. https://ourworldindata.org/natural-disasters.

[Ritchie, Roser, and Ortiz-Ospina 2015] Ritchie, H.; Roser, M.; and Ortiz-Ospina, E. 2015. Suicide. *Our World in Data*. https://ourworldindata.org/suicide.

[Roth et al. 2018] Roth, G. A.; Abate, D.; Abate, K. H.; Abay, S. M.; Abbafati, C.; Abbasi, N.; Abbastabar, H.; Abd-Allah, F.; Abdela, J.; Abdelalim, A.; et al. 2018. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet* 392(10159):1736–1788.

[Stevens and Sogolow 2005] Stevens, J. A., and Sogolow, E. D. 2005. Gender differences for non-fatal unintentional fall related injuries among older adults. *Injury prevention* 11(2):115–119.

[Ueno et al. 2019] Ueno, F.; Tamaki, R.; Saito, M.; Okamoto, M.; Saito-Obata, M.; Kamigaki, T.; Suzuki, A.; Segubre-Mercado, E.; Aloyon, H. D.; Tallo, V.; Lupisan, S. P.; Oshitani, H.; and in the Philippines, R. W. G. 2019. Age-specific incidence rates and risk factors for respiratory syncytial virus-associated lower respiratory tract illness in cohort children under 5 years old in the philippines. *Influenza and Other Respiratory Viruses* 13(4):339–353.

[White et al. 2014] White, M. C.; Holman, D. M.; Boehm, J. E.; Peipins, L. A.; Grossman, M.; and Henley, S. J. 2014. Age and cancer risk: a potentially modifiable relationship. *American journal of preventive medicine* 46(3):S7–S15.