FAIKR-Mod3 Project Reports

Lorenzo Cassano, Jacopo D'Abramo
Master's Degree in Artificial Intelligence, University of Bologna
lorenzo.cassano2@studio.unibo.it, jacopo.dabramo@studio.unibo.it

October 16, 2022

Abstract

The project is based on a dataset which represents the job situation of an employee and the main focus was to study the difference between a Bayesian Network obtained by structure learning and a handmade network. First of all, to obtain the best network, structure learning algorithms were used and then the net was compared to an hand-made one. In light of above, several experiments has been done using both exact and approximate inference: in the exact inference the work focused on running queries for different reasoning pattern and in approximate it was tested time performance in different types of sampling. To put it in briefly, in this project has been used Belief Networks to handle uncertainty and it has been done different experiments. Basic knowledge in probability theory are essential to understanding the content of this report.

Introduction

Domain

Our main idea was to define a Belief Network in order to have a conditional probability causal-effect.Thus, starting from the scientific article: "Performance measures and worker productivity" (J.), we have found a dataset in the domain of employment attrition which appeared very good for our case of study. In particular, dataset's features showed a correlation between personal and job information dealing with our aim to build causal-effect Belief Network. For i.e, observing the domain the attrition may depend on: age, work overtime and working hours of employee; as long as salary may be conditioned from working position, seniority of employee and work overtime.

Aim

The purposes which involved us in this projects were many. In particular we focused in different areas of building and making inference in a belief network. In the following points we summarize the main objectives:

- trying different methods to build a belief network:first using structure learning techniques proposed by pgmpy library and then defining a hand-made network based on our idea of dependencies. Comparing the models obtained and their complexity.
- Studying different types of parameters and their impact on the net both in building and in inference
- Analyzing the dependencies and indipendence of the networks among random variables and how they change on the base of specific evidences.
- make inference using both the methods studied during the lectures (exact and approximate) comparing and analyzing the differences and similarities.

Method

The methodology followed can be divided in three main categories: Building, Analyzing and making Inference. All of the experiments performed has been run on two different types of network: one obtained from structure learning techniques and the other designed by hand. The main tool which we used is (Ankan and Panda 2015) pgmpy a open-source python library for designing belief network.

- Building: Concerning the learned network we approached to the main algorithms proposed by the aforementioned library. The first one proposed is PC (Ankan and Panda 2015)(constraint-based Estimator) algorithm which did not lead to good results so the next estimator involved has been (Russell 2010) Hill Climb search which brought a large improvements working on implicit parameters (BicScore,K2Score (Ankan and Panda 2015)) and on edges . On the other hand, the hand made net has been designed according to the structural equation principle (Russell 2010).
- Analyzing: The main topic covered in our analysis are: complexity and indipendence. The former has been studied analyzing the number of indipendence parameters of the network whereas the latter computing active trail between two random variables. Moreover in both of the net the calculution of the CPT's has been done using a BayesianEstimator due to 0-probability problem of data.
- Inference: Both approximate and exact inference have been studied for the networks. Specifically in exact inference we computed three types of reasoning patterns : Causal, Evidential and Intercausal (Koller and Friedman 2009). By way of an alternative in approximate inference the time complexity has been brought on.

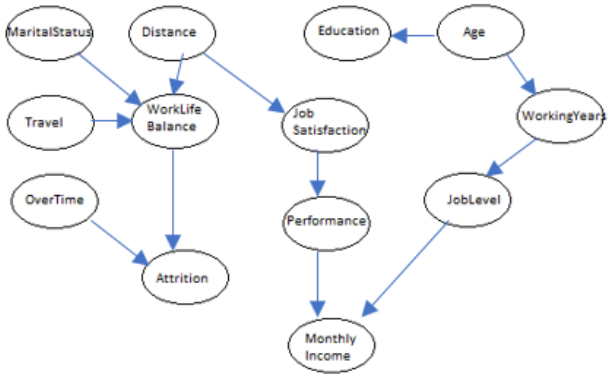

Figure 1: Learned Bayesian Network

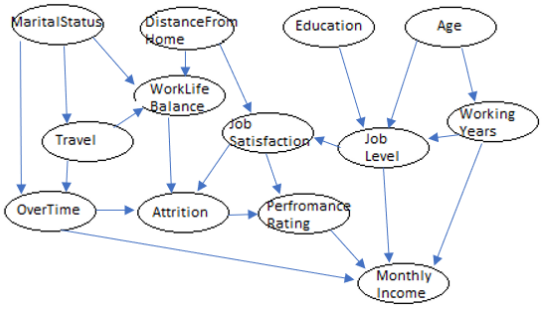

Figure 2: Hand made bayesian network

Results

The most interesting result regards the complexity of the networks obtained and all the aspect of inference from the probabilities queries to the differences of the sampling algorithm.

Model

To begin with bayesian net, self-explained in Figure 1, the procedure followed the application of structure learning algorithms in order to get the model which fit the data in the best way. Therefore we reached a good compromise with the usage of the Hill Climbing searching method combined with a fixed and black list of edges. The black list has been designed such that an edge from a r.v. X to a r.v. Y exists if and only if X comes before Y in the sort of r.v.. Whereas the fixed edges list contains edges that the algorithm has not been able to discover but necessary in the domain. The handmade net (Figure 2) basically has been built respecting the aforementioned structural equation and so a r.v X depends from a r.v. Y if and only if the nature assigns a truth value to X based on what learns about Y. So we followed a causal reasoning starting from the caused to the effects.

Analysis

Experimental setup

In the first place we experimented how complex the two nets are. By doing this we determined the number of independent parameters and the bigger is the number the more complex is the net. According to the networks we expected the hand made to be more convoluted. Concerning inference, we run three different kinds of probability queries, below an example for each category:

- Causal Inference:
  $P(MonthlyIncome \mid TotalWorkingYears=Proficient,\ JobSatisfaction = High\ )$
- Evidencial Inference:
  $P(Age \mid MonthlyIncome = Higher\ )$
- Intercausal Inference:
  $P(JobLevel \mid MonthlyIncome = Lower,\ PerformanceRating = Outstanding)$

For approximate inference, we performed sampling with five different size *[10,100,500,number of records, 1550]* and for each the two main algorithms considered both the presence and absences of evidences.

Results

According to our expectations, the network obtained from structure learning algorithms has more indipendence and less complex as it is possible observe during conditional independent analysis. Other than that, the probabilities got from the queries are very similar in both networks with respect to our supposition. Concerning the time performance analysis, on the two different sampling, what came up in both networks is that the time of likelihood weighted sample remains equal whether we use observe variables or not, on the other hand, in rejection sampling there are great differences in time sampling using observed variable or not.

Conclusion

The models obtained differ in complexity and detailing and so it is possible to expand the inference in the hand net with specific queries. To sum up this project allowed us to deepen the topics of structure learning techniques which lead to interesting results and at the same time to put into practice the methods and algorithms studied during the lectures.

Links to external resources

- GitHub repository `https://github.com/jacopodabramo/FAIKR3`
- Scientific paper `https://wol.iza.org/articles/performance-measures-and-worker-productivity`

References

Ankan, A., and Panda, A. 2015. pgmpy: Probabilistic graphical models using python.

J., S. Performance measures and worker productivity. *IZA World of Labor* 260.

Koller, D., and Friedman, N. 2009. Probabilistic graphical models: Principles and techniques.

Russell, S. J. 2010. *Artificial intelligence a modern approach*. Pearson Education, Inc.