Bayesian Network Model for Analyses of Road Accidents

Fundamentals of Artificial Intelligence and Knowledge Representation (module 3)

Francesca Boccardi, francesca.boccardi@studio.unibo.it

Luigi Podda, luigi.podda@studio.unibo.it

Abstract

Road traffic injuries represent an important public health problem. Preventive strategies are one of the most effective methods in reducing the number of road accidents and, as a consequence, of injuries. For the purpose of analyzing road accidents causes, the Bayesian network is a powerful tool, able to predict the probability of injuries in certain road traffic conditions and to find the critical states combination which may leads to unsafe situations. In this study, starting from a suitable data set, a model for road accidents has been formulated using a Bayesian network. Then, some inferential analyses on accident severities have been conducted, using both exact and approximate inference methods. Finally, the found results have been plotted and summarized.

Introduction

A Bayesian network (BN) is a probabilistic graphical model for representing knowledge about an uncertain domain where each node corresponds to a random variable and each edge represents the conditional probability for the corresponding random variables. Due to dependencies and conditional probabilities, a BN corresponds to a directed acyclic graph (DAG) where no loop or self connection is allowed.[1]
Starting from this definition, we modelled a Bayesian network able to represent the road injuries context for the purpose of deducing the effects of an accident from its causes. This analysis is done to sensitize people on the consequences that their behavior on the road can conduct, but also to inform them that there exist circumstances beyond their control where a careful manner can importantly influence the event's results.

Dataset and Preprocessing

The preliminary step to model a Bayesian network regards the data preprocessing.
Starting from two complex datasets[2], respectively containing information about the accidents and information about vehicles and drivers, a more compact one is created by the merging of the two. After this first process, some further manipulations have been added:

1. dropping of the unused columns;
2. discretization of continuous features;
3. quantization of some features in order to obtain a more compact data representation;
4. merging of some features in order to obtain a more synthetic data representation.

Finally, the resulting dataset counts 9 features:

Table 1: Dataset features description

FEATURE	VALUES	DESCRIPTION
Age	Age group bins	User's age
Speed	*under speed limits*, *over speed limits*	User's speed at accident time
RoadUserConditions	*Not Altered*, *Under Drug Influence*, *Had Been Drinking*, *Sleepy/Fatigued*, *Impairment Physical*	User's physical conditions
PartyType	*Vulnerable Users*, *Car*, *Truck*, *Motorcycle*, *Train*, *Bus*	Involved vehicle/user type
CrashTime	Time slots	Crash time
RoadConditions	*Dry*, *Wet*, *Slippery*	Road conditions
Lighting	*Dark - Street Light*, *Dark - No Street Light*, *Daylight*, *Dusk - Dawn*	Lighting road conditions
Weather	*Clear*, *Cloudy*, *Wind*, *Rain*, *Fog*	Weather conditions
Injuries	*0* or *1*	0 if the road accident caused no or only minor injuries, 1 if it caused severe or fatal injuries

Bayesian Network construction

We used the *pgmpy* library to create the Bayesian network, specifying the nodes and their relationship with the others. In particular, the set of nodes represents the features (columns) of our dataset and each edge represents the conditional probability for the nodes it links.

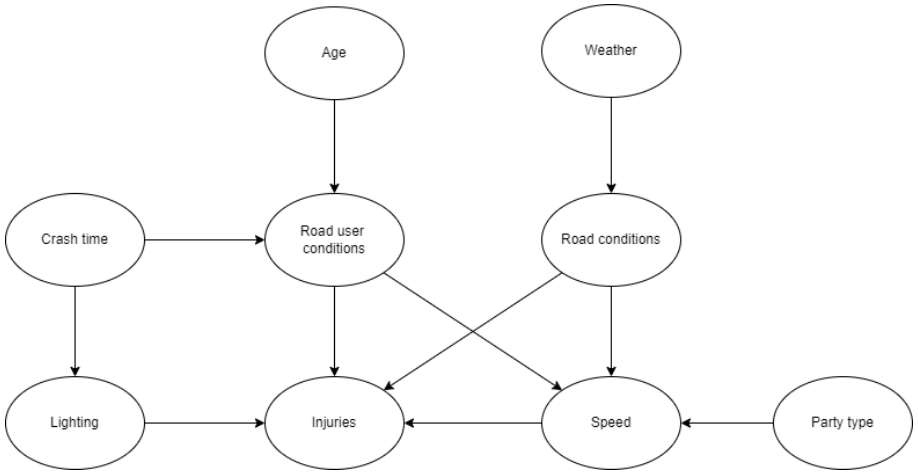

Figure 1: Bayesian network

Model description

In this section we explain our network design choices, considering that the Bayesian network follows cause-effect relationships:

- **Road User Conditions** → the behavior of the user strongly depends on its age, as the maturity may affect the carefulness of the individual in different ways. Moreover, the lucidity of a person can be really influenced by the day time. Thus, *Crash time* and *Age* have been modelled to be *Road user conditions* parents.
- **Lighting** → lighting conditions depend on the crash time, as they are directly affected by the day time.
- **Road conditions** → road states are directly influenced by the weather conditions.
- **Speed** → the speed is affected by multiple factors: road conditions, vehicle/user involved in the accident (party type) and road user's conditions, which in particular describes the general lucidity of the user.
- **Injuries** → several aspects influence whether or not an accident causes injuries: the lighting conditions (if the road is lighted up or not), the conditions of the user (lucidity), the road state and the user's speed (if it is under or over the allowed limits).
  Indeed, the combination of particular instances of these variables can lead to different levels of danger, which can determine the seriousness of the accident.

Conditional Probability Distributions

Once the structure of the network has been defined, we fit our model on the preprocessed dataset using the Maximum Likelihood Estimator, in order to compute the **CPD** table for all the features.
Given the theoretical concepts related to the Bayesian networks, it turned out to be useful to run some experiments on the model, in order to acquire a deeper knowledge of it. In particular, to better analyze the properties of the network, we chose to explore and visualized its relative **local independencies**, the **Markov Blanket** of each node and some of the most interesting **active trails** between its features.

Inference

Bayesian networks are really powerful tools for operating reasoning processes on variables.
In particular, probabilistic inference on unobserved variables, called *query variables*, can be done following different reasoning patterns, which depend on the direction of the information propagation:

- *Causal reasoning* → given evidences on causes, can be made predictions on the downstream effects
- *Evidential reasoning* → given evidences on effects, can be computed the likelihood that any of their causes was the contributing factor
- *Intercausal reasoning* → given evidences on a cause and on its effects, can be deduced information about the other causes

With the aim of inferring some useful information about the potential accident severities in particular traffic conditions, we asked our model several probabilistic queries, using both exact and approximate inference methods, eventually comparing the found results.

Exact Inference

In exact inference, the conditional probability distribution over the variables of interest is computed analytically from the CPD tables. To do so, we used the **variable elimination** method.

Remarkable queries

In this section, we will show the results of two of the most interesting queries, computed by following the causal reasoning.

1. Given that a road user is under drug influence, what's the probability that, in the case of an accident, this will cause injuries?

The apriori probability distribution of an accident to cause *Injuries* is:

Table 2: P(Injuries)

Injuries	$\Phi(Injuries)$
0	0.6968
1	0.3032

The conditional probability distribution of an accident to cause *Injuries* given as evidence *RoadUserConditions = Under Drug Influence* is:

Table 3: P(Injuries | RoadUserConditions = Under Drug Influence)

Injuries	$\Phi(Injuries)$
0	0.6204
1	0.3796

Can be clearly noticed that the probability of a potential road accident to cause injuries increases if the road user is under drug influence.

2. Given that a road user had been drinking, what's the probability that, in the case of an accident, its speed will be over the allowed limits?

The apriori probability distribution of the user's *Speed* at accident time is:

Table 4: P(Speed)

Speed	$\Phi(Speed)$
over speed limits	0.1368
under speed limits	0.8632

The conditional probability distribution of the user's *Speed* at accident time, given as evidence *RoadUserConditions = Had Been Drinking* is:

Table 5: P(Speed | RoadUserConditions = Had Been Drinking)

Speed	$\Phi(Speed)$
over speed limits	0.2696
under speed limits	0.7304

Can be clearly noticed that, in the case of an accident, the probability of the road user's speed to be over the allowed limits increases if the user had been drinking.

Approximate Inference

Approximate inference is a method for estimate probabilities in Bayesian networks. The usage of approximate inference is crucial when computing posterior probabilities with exact inference becomes intractable, as for the case of large multiply-connected networks, where the variable elimination method can have exponential time and space complexity.
There exist different methods to compute the probabilities using approximate inference, but in this study we applied just two of them: the **Rejection sampling** and the **Likelihood weighted sampling**.
In particular, we focused our analysis on computing the conditional probability of

$$P(Injuries = 1 \mid RoadUserConditions = Under\ Drug\ Influence)$$

with both the exact and approximate inference, in order to compare their relative outcomes.

Results and plots

As for the computations performed using the exact inference, the results show that the probability of a potential road accident to cause injuries, given that the road user is under drug influence, is about 38%.

$$P(Injuries = 1 \mid RoadUserConditions = Under\ Drug\ Influence) = 0.3795950064306464$$

As for the probability estimation with approximate inference, the run of several experiments with both rejection sampling and likelihood weighted methods gave good results exclusively when the sample size was relatively large.

To better visualize the comparison between these inference techniques, it's helpful to plot the results.

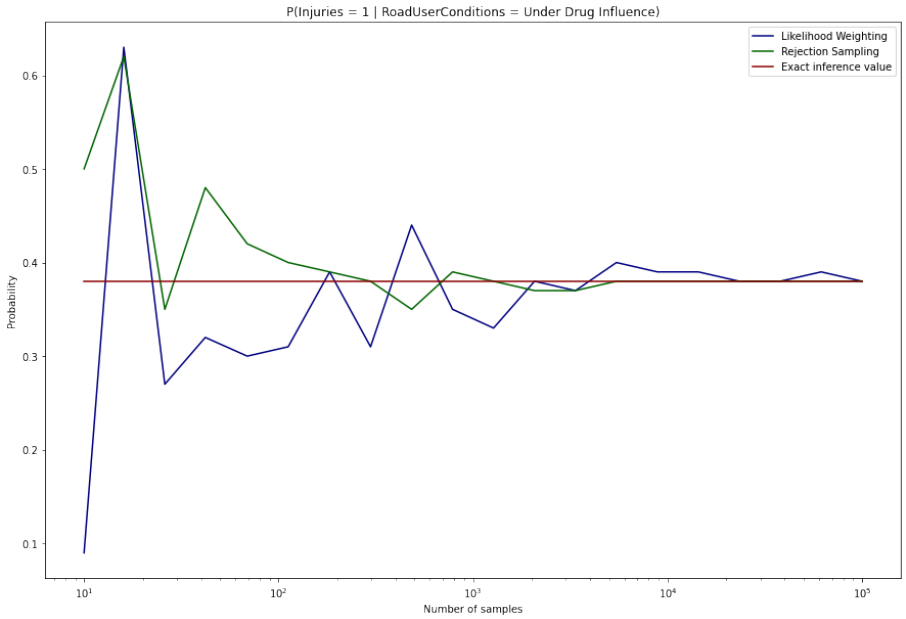

Figure 2: Probabilities

In addition, a further analysis on the absolute errors of the two approximated probabilities with respect to the exact inference can be done.

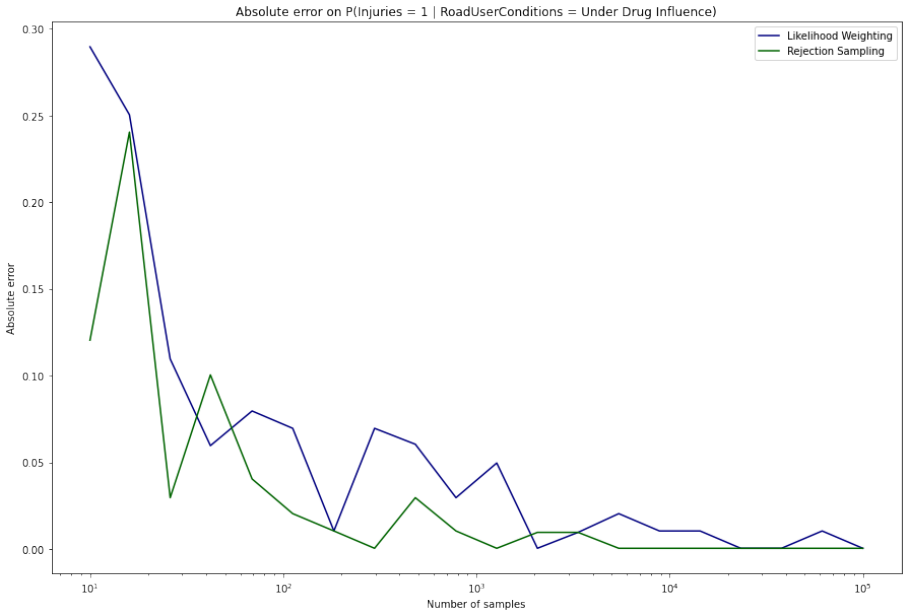

Figure 3: Absolute Error

Finally, can be definitely observed that the higher the number of samples, the better the approximation of the probability to be estimated. Indeed, with a suitable number of samples, both the two approximate methods converge to the value of the exact inference probability.

Conclusions

Bayesian networks are really useful tool for dealing with probabilistic systems, mainly thanks to the fully explainability of their reasoning process.

During this study, due to the lack of appropriate and complete datasets, we were forced to use a restricted set of data and to adapt on it the whole model. Nevertheless, the developed network turned out to be comprehensive enough to allow the conduction of satisfactory inferential analyses.

References

[1] Xin-She Yang, *Introduction to Algorithms for Data Mining and Machine Learning*, 2019.

[2] Dataset source: `https://data.sanjoseca.gov/dataset/crashes-data`.

[3] Xin Zou and Wen Long Yue, *A Bayesian Network Approach to Causation Analysis of Road Accidents Using Netica*, 2007 (`https://www.hindawi.com/journals/jat/2017/2525481/`).

[4] Bayesian network and reasoning patterns: `https://classes.engr.oregonstate.edu/eecs/winter2020/cs536/slides/bayesnets2.2pp.pdf`.

[5] Judea Perl, Stuart Russel, *Bayesian Networks*, 2000, (`https://escholarship.org/content/qt53n4f34m/qt53n4f34m.pdf`).