BayesSuites: An Open Web Framework for Visualization of Massive Bayesian Networks

Nikolas Bernaola   N.BERNAOLA@ALUMNOS.UPM.ES

Mario Michiels

Concha Bielza

Pedro Larrañaga
Computational Intelligence Group, Departamento de Inteligencia Artificial, ETSI Informáticos, Universidad Politécnica de Madrid

Abstract

BayesSuites [1] is the first web framework for learning, visualizing, and interpreting Bayesian networks that can scale to tens of thousands of nodes while providing fast and friendly user experience. BayesSuites solves the problems of scalability, extensibility and interpretability that massive networks bring by separating backend calculations from the frontend interface and using specialized learning algorithms for massive networks. We demonstrate the tool by learning and visualizing a genome-wide gene regulatory network from human brain data with 20,708 nodes.

Keywords: Bayesian networks; Interpretability; Visualization of massive networks; Gene regulatory networks

1. Introduction

In recent years, there has been an explosion in the size and complexity of data being generated that has to be analyzed to extract meaningful information. This data explosion is occurring simultaneously in many fields as new data acquisition technologies become available and data sharing becomes the norm. Among the techniques used to interpret and analyze these datasets, machine learning, specially using algorithms such as neural networks, has become more widely used. Among the main limitations of most of the commonly used machine learning techniques is the black box problem, where there is no explanation as to why the model has given one output over another for a given input.

Here we will focus on interpretable AI which abandons the use of black box models in favour of models that are inherently legible and transparent to the user. The requirements for a model to be interpretable are: (a) conclusion transparency; it should be possible to ascertain why the conclusion was reached by showing the relationships between the variables and (b) improvement of the system; it should be possible to not only validate the results but to change the model in a way that allows the user to see how the changes affect the conclusion. In general, both of these will be hard even in transparent models when we are dealing with very complex systems (the relationships might be clear, but checking all of them might take too long) so we need better tools to aid us in the task of reading these systems.

In this work we focus on Bayesian networks (BNs) Koller and Friedman (2009) Maathuis et al. (2018), a class of probabilistic graphical models that fulfill the interpretability requirements stated

---

1. BayesSuites webpage

above in an intuitive way since the relationships between variables can be directly interpreted in terms of probabilistic conditional independencies. The inherently graphical nature of this family of models makes it very easy to visualize them and the inference machinery allows for changes in the variables and to easily draw conclusions on what those changes imply. Although theoretically a very good choice for interpretable systems, in practice BNs suffer from many different problems that make them be relegated in many use cases.

The rest of this manuscript focuses on better explaining these limitations, how our framework BayesSuites included in the NeuroSuties platform shown in Michiels et al. (2020), tries to solve them and concludes with an application of our software to one domain where high-dimensional datasets abound as is the construction of genome-wide gene regulatory networks (GRNs) as explained in Bernaola et al. (2020).

2. BayesSuites

2.1 Scalability

Massive BNs present mainly three scalability problems: learning their graph structure, efficiently visualizing it, and developing a fast inference engine for the probabilistic reasoning. BayesSuites solves these issues by being implemented as a scalable web application that runs the heavy operations in the backend while providing a lightweight and fast frontend.
Of the multiple learning algorithms available in the platform some are suitable for large-scale networks, but only FGES-Merge works in a reasonable time without constraining the structure. This algorithm works through a parallelized greedy search in which small neighbourhood networks are learnt around each node and then merged together and pruned to obtain the full network.
As for the inference, it also runs in the backend. For the use case of GRNs we use Gaussian nodes which massively increases the speed of inference compared with more general distributions.

2.2 Extensibility

BayesSuites is built in a modular way and is easy to extend by, for example, adding new learning algorithms. The backend is written in Python, so the easiest way to extend would be to code a Python module. However, we aimed to support as many scientific users as possible so there is also a way to extend BayesSuites using R by through the Rpy2 package.

2.3 Ease of Use and Interpretability

We review the capabilities of BayesSuites by presenting a real use case: learning and interpreting the GRN of the full human genome using microarray data from the Allen Brain Atlas. The dataset contains 20,708 protein-coding genes each of which correspond to one of the features of the network.
The dataset can be easily uploaded and, then, a structure learning method is chosen from the ones available. In this case, we used FGES-Merge. Once the structure is finished, we learn the parameters and show the network to the user.
From here, we have the many layout options as shown in Fig. 2 to change many of the parameters of the network and zoom on relevant variables. A final global view of the network might look like Fig. 1. Then we could zoom on a variable of interest and perform inference only on it and its Markov

Figure 1: Example of a genome-wide regulatory network learnt with human brain data. Shown groups were colored using the Louvain Blondel et al. (2008) algorithm. The network was learnt using FGES-Merge. Image from Bernaola et al. (2020) taken with permission of the authors.

blanket or use groups to keep track of different variables of interest and how they relate to each other.

3. Conclusion

BayesSuites is a software tool that has been designed to make it easy to learn, visualize and operate with massive Bayesian networks so that one of the biggest advantages of graphical models, which is their interpretability, is not lost due to the sheer complexity of navigating a network.

References

N. Bernaola, M. Michiels, P. Larrañaga, and C. Bielza. Learning massive interpretable gene regulatory networks of the human brain by merging Bayesian networks. *BioRxiv*, 2020.

V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

M. Maathuis, M. Drton, S. Lauritzen, and M. Wainwright. *Handbook of graphical models*. CRC Press, 2018.

M. Michiels, P. Larrañaga, and C. Bielza. Neurosuites-BNs: An open web framework for massive Bayesian networks focused on neuroscience. *bioRxiv*, 2020.
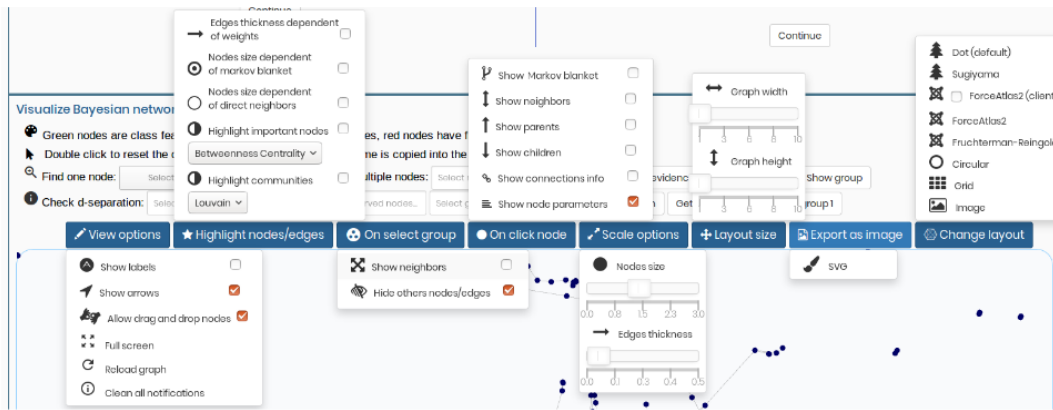
(a) Lower bar options
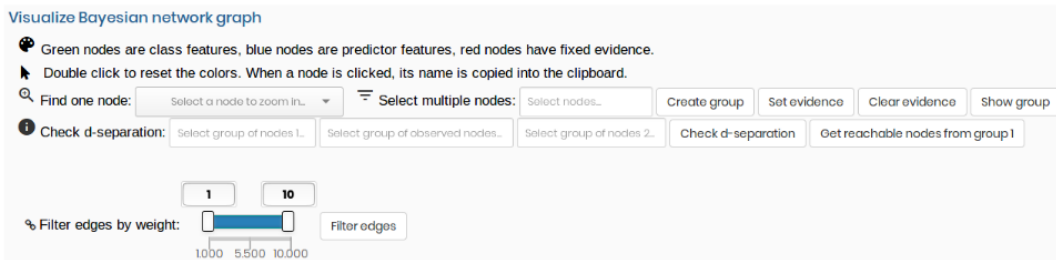
(b) Upper bar options
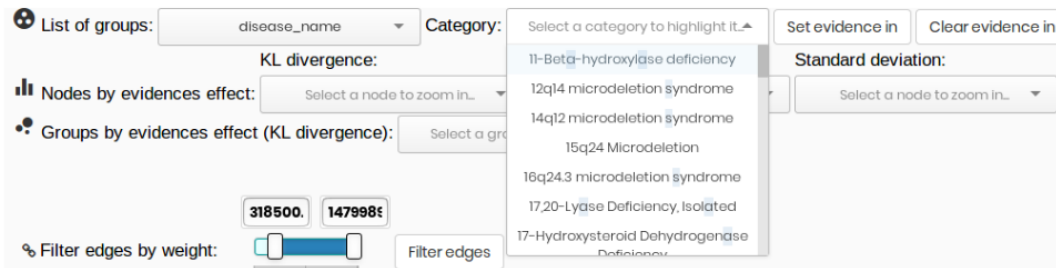
(c) Lower bar options when groups are created and inference is conducted

Figure 2: Layout options in BayesSuites. (a) Shows the visualization options, with the network layouts, edge and node sizes and colours, and highlighting of parents, children or Markov blankets. (b) Shows the options to create groups and filter nodes by degree or by metadata (which can be included when uploading the dataset). (c) Shows inference options for groups. Instead of setting evidence for just one variable, all of the group can be set to some value (absolute or as a multiple of the standard deviation) and shows the options to visualize effect sizes of the introduced evidence in the chosen nodes. Image from Michiels et al. (2020) taken with permission of the authors.