Learning Bayesian Networks

Marco F. Ramoni
Harvard Medical School, USA

Paola Sebastiani
Boston University School of Public Health, USA

INTRODUCTION

Born at the intersection of artificial intelligence, statistics, and probability, Bayesian networks (Pearl, 1988) are a representation formalism at the cutting edge of knowledge discovery and data mining (Heckerman, 1997). Bayesian networks belong to a more general class of models called *probabilistic graphical models* (Whittaker, 1990; Lauritzen, 1996) that arise from the combination of graph theory and probability theory, and their success rests on their ability to handle complex probabilistic models by decomposing them into smaller, amenable components. A probabilistic graphical model is defined by a graph, where nodes represent stochastic variables and arcs represent dependencies among such variables. These arcs are annotated by probability distribution shaping the interaction between the linked variables. A probabilistic graphical model is called a Bayesian network, when the graph connecting its variables is a directed acyclic graph (DAG). This graph represents conditional independence assumptions that are used to factorize the joint probability distribution of the network variables, thus making the process of learning from a large database amenable to computations. A Bayesian network induced from data can be used to investigate distant relationships between variables, as well as making prediction and explanation, by computing the conditional probability distribution of one variable, given the values of some others.

BACKGROUND

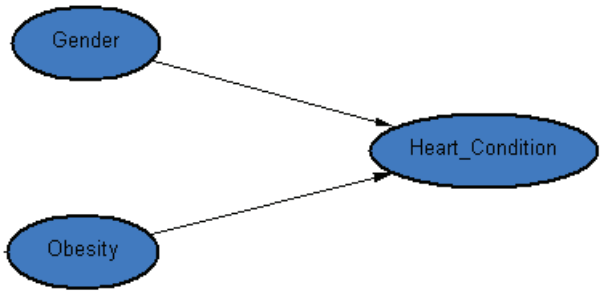
The origins of Bayesian networks can be traced back as far as the early decades of the 20th century, when Sewell Wright developed path analysis to aid the study of genetic inheritance (Wright, 1923, 1934). In their current form, Bayesian networks were introduced in the early 1980s as a knowledge representation formalism to encode and use the information acquired from human experts in automated reasoning systems in order to perform diagnostic, predictive, and explanatory tasks (Charniak, 1991; Pearl, 1986, 1988). Their intuitive graphical nature and their principled probabilistic foundations were very attractive features to acquire and represent information burdened by uncertainty. The development of amenable algorithms to propagate probabilistic information through the graph (Lauritzen, 1988; Pearl, 1988) put Bayesian networks at the forefront of artificial intelligence research. Around the same time, the machine-learning community came to the realization that the sound probabilistic nature of Bayesian networks provided straightforward ways to learn them from data. As Bayesian networks encode assumptions of conditional independence, the first machine-learning approaches to Bayesian networks consisted of searching for conditional independence structures in the data and encoding them as a Bayesian network (Glymour, 1987; Pearl, 1988). Shortly thereafter, Cooper and Herskovitz (1992) introduced a Bayesian method that was further refined by Heckerman, et al. (1995) to learn Bayesian networks from data.

These results spurred the interest of the data-mining and knowledge-discovery community in the unique features of Bayesian networks (Heckerman, 1997); that is, a highly symbolic formalism, originally developed to be used and understood by humans, well-grounded on the sound foundations of statistics and probability theory, able to capture complex interaction mechanisms and to perform prediction and classification.

MAIN THRUST

A Bayesian network is a graph, where nodes represent stochastic variables and (arrowhead) arcs represent

Figure 1.



dependencies among these variables. In the simplest case, variables are discrete, and each variable can take a finite set of values.

Representation

Suppose we want to represent the variable *gender*. The variable gender may take two possible values: male and female. The assignment of a value to a variable is called the *state of the variable*. So, the variable gender has two states: Gender = Male and Gender = Female. The graphical structure of a Bayesian network looks like this:

The network represents the notion that obesity and gender affect the heart condition of a patient. The variable obesity can take three values: yes, borderline and no. The variable heart condition has two states: true and false. In this representation, the node heart condition is said to be a *child* of the nodes gender and obesity, which, in turn, are the *parents* of heart condition.

The variables used in a Bayesian networks are stochastic, meaning that the assignment of a value to a variable is represented by a probability distribution. For instance, if we do not know for sure the gender of a patient, we may want to encode the information so that we have better chances of having a female patient rather than a male one. This guess, for instance, could be based on statistical considerations of a particular population, but this may not be our unique source of information. So, for the sake of this example, let's say that there is an 80% chance of being female and a 20% chance of being male. Similarly, we can encode that the incidence of obesity is 10%, and 20% are borderline cases. The following set of distributions tries to encode the fact that obesity increases the cardiac risk of a patient, but this effect is more significant in men than women:

The dependency is modeled by a set of probability distributions, one for each combination of states of the variables gender and obesity, called the parent variables of heart condition.

Figure 2.

Heart_Condition

Obesity	Gender	True	False
Yes	Male	0.800	0.200
Yes	Female	0.700	0.300
Borderline	Male	0.750	0.250
Borderline	Female	0.600	0.400
No	Male	0.200	0.800
No	Female	0.100	0.900

Learning

Learning a Bayesian network from data consists of the induction of its two different components: (1) the graphical structure of conditional dependencies (model selection) and (2) the conditional distributions quantifying the dependency structure (parameter estimation).

There are two main approaches to learning Bayesian networks from data. The first approach, known as constraint-based approach, is based on conditional independence tests. As the network encodes assumptions of conditional independence, along this approach we need to identify conditional independence constraints in the data by testing and then encoding them into a Bayesian network (Glymour, 1987; Pearl, 1988; Whittaker, 1990).

The second approach is Bayesian (Cooper & Herskovitz, 1992; Heckerman et al., 1995) and regards model selection as an hypothesis testing problem. In this approach, we suppose to have a set M= $\{M_0, M_1, ..., M_g\}$ of Bayesian networks for the random variables $Y_1, ..., Y_v$,, and each Bayesian network represents an hypothesis on the dependency structure relating these variables. Then, we choose one Bayesian network after observing a sample of data D = $\{y_{1k}, ..., y_{vk}\}$, for $k = 1, . . . , n$. If $p(M_h)$ is the prior probability of model M_h, a Bayesian solution to the model selection problem consists of choosing the network with maximum posterior probability:

$p(M_h|D) \propto p(M_h)p(D|M_h)$.

The quantity $p(M_h|D)$ is the marginal likelihood, and its computation requires the specification of a parameterization of each model M_h and the elicitation of a prior distribution for model parameters. When all variables are discrete or all variables are continuous, follow Gaussian distributions, and the dependencies are linear and the marginal likelihood factorizes into the product of marginal likelihoods of each node and its parents. An important property of this likelihood modularity is that in the comparison of models that differ only for the parent structure of a variable Y_i, only the local marginal likelihood matters. Thus, the comparison of two local network structures that specify different parents for Y_i can be done simply by evaluating the product of the local Bayes factor $BF_{h,k} = p(D|M_{hi}) / p(D|M_{ki})$, and the ratio $p(M_{hi})/ p(M_{ki})$, to compute the posterior odds of one model vs. the other as $p(M_{hi}|D) / p(M_{ki}|D)$.

In this way, we can learn a model locally by maximizing the marginal likelihood node by node. Still, the space of the possible sets of parents for each variable grows exponentially with the number of parents involved, but successful heuristic search procedures (both deterministic and stochastic) exist to render the task more amenable (Cooper & Herskovitz, 1992; Singh & Larranaga, 1996; Valtorta, 1995).

Once the structure has been learned from a dataset, we still need to estimate the conditional probability distributions associated to each dependency in order to turn the graphical model into a Bayesian network. This process, called *parameter estimation*, takes a graphical structure and estimates the conditional probability distributions of each parent-child combination. When all the parent variables are discrete, we need to compute the conditional probability distribution of the child variable, given each combination of states of its parent variables. These conditional distributions can be estimated either as relative frequencies of cases or, in a Bayesian fashion, by using these relative frequencies to update some, possibly uniform, prior distribution. A more detailed description of these estimation procedures for both discrete and continuous cases is available in Ramoni and Sebastiani (2003).

Prediction and Classification

Once a Bayesian network has been defined, either by hand or by an automated discovery process from data, it can be used to reason about new problems for prediction, diagnosis, and classification. Bayes' theorem is at the heart of the propagation process.

One of the most useful properties of a Bayesian network is the ability to propagate evidence irrespective of the position of a node in the network, contrary to standard classification methods. In a typical classification system, for instance, the variable to predict (i.e., the class) must be chosen in advance before learning the classifier. Information about single individuals then will be entered, and the classifier will predict the class (and only the class) of these individuals. In a Bayesian network, on the other hand, the information about a single individual will be propagated in any direction in the network so that the variable(s) to predict must not be chosen in advance.

Although the problem of propagating probabilistic information in Bayesian networks is known to be, in the general case, NP-complete (Cooper, 1990), several scalable algorithms exist to perform this task in networks with hundreds of nodes (Castillo, et al., 1996; Cowell et al., 1999; Pearl, 1988). Some of these propagation algorithms have been extended, with some restriction or approximations, to networks containing continuous variables (Cowell et al., 1999).

FUTURE TRENDS

The technical challenges of current research in Bayesian networks are focused mostly on overcoming their current limitations. Established methods to learn Bayesian networks from data work under the assumption that each variable is either discrete or normally distributed around a mean that linearly depends on its parent variables. The latter networks are termed *linear Gaussian* networks, which still enjoy the decomposability properties of the marginal likelihood. Imposing the assumption that continuous variables follow linear Gaussian distributions and that discrete variables only can be parent nodes in the network but cannot be children of any continuous node, leads to a closed-form solution for the computation of the marginal likelihood (Lauritzen, 1992). The second technical challenge is

the identification of sound methods to handle incomplete information, either in the form of missing data (Sebastiani & Ramoni, 2001) or completely unobserved variables (Binder et al., 1997). A third important area of development is the extension of Bayesian networks to represent dynamic processes (Ghahramani, 1998) and to decode control mechanisms.

The most fundamental challenge of Bayesian networks today, however, is the full deployment of their potential in groundbreaking applications and their establishment as a routine analytical technique in science and engineering. Bayesian networks are becoming increasingly popular in various fields of genomic and computational biology—from gene expression analysis (Friedman, 2004) to proteimics (Jansen et al., 2003) and genetic analysis (Lauritzen & Sheehan, 2004)—but they are still far from being a received approach in these areas. Still, these areas of application hold the promise of turning Bayesian networks into a common tool of statistical data analysis.

CONCLUSION

Bayesian networks are a representation formalism born at the intersection of statistics and artificial intelligence. Thanks to their solid statistical foundations, they have been turned successfully into a powerful data-mining and knowledge-discovery tool that is able to uncover complex models of interactions from large databases. Their high symbolic nature makes them easily understandable to human operators. Contrary to standard classification methods, Bayesian networks do not require the preliminary identification of an outcome variable of interest, but they are able to draw probabilistic inferences on any variable in the database. Notwithstanding these attractive properties and the continuous interest of the data-mining and knowledge-discovery community, Bayesian networks still are not playing a routine role in the practice of science and engineering.

REFERENCES

Binder, J. et al. (1997). Adaptive probabilistic networks with hidden variables. *Mach Learn, 29*(2-3), 213-244.

Castillo, E. et al. (1996). *Expert systems and probabiistic network models*. New York: Springer.

Charniak, E. (1991). Bayesian networks without tears. *AI Magazine, 12*(8), 50-63.

Cooper, G.F. (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artif Intell, 42*(2-3), 393-405.

Cooper, G.F., & Herskovitz, G.F. (1992). A Bayesian method for the induction of probabilistic networks from data. *Mach Learn, 9*, 309-347.

Cowell, R.G., et al. (1999). *Probabilistic networks and expert systems*. New York: Springer.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science, 303*, 799-805.

Ghahramani, Z. (1998). Learning dynamic Bayesian networks. In C.L. Giles, & M. Gori (Eds.), *Adaptive processing of sequences and data structures* (pp. 168-197). New York: Springer.

Glymour, C., Scheines, R., Spirtes, P., & Kelly, K. (1987). *Discovering causal structure: Artificial intelligence, philosophy of science, and statistical modeling*. San Diego, CA: Academic Press.

Heckerman, D. (1997). Bayesian networks for data mining. *Data Mining and Knowledge Discovery, 1*(1), 79-119.

Heckerman, D. et al. (1995). Learning Bayesian networks: The combinations of knowledge and statistical data. *Mach Learn, 20*, 197-243.

Jansen, R. et al. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science, 302*, 449-453.

Larranaga, P., Kuijpers, C., Murga, R., & Yurramendi, Y. (1996). Learning Bayesian network structures by searching for the best ordering with genetic algorithms. *IEEE T Syst Man Cyb, 26*, 487-493.

Lauritzen, S.L. (1992). Propagation of probabilities, means and variances in mixed graphical association models. *J Amer Statist Assoc, 87*, 1098-108.

Lauritzen, S.L. (1996). *Graphical models*. Oxford: Clarendon Press.

Lauritzen, S.L. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J Roy Stat Soc B Met, 50*, 157-224.

Lauritzen, S.L., & Sheehan, N.A. (2004). Graphical models for genetic analysis. *Statist Sci, 18*(4), 489-514.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artif. Intell., 29*(3), 241-288.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco: Morgan Kaufmann.

Ramoni, M., & Sebastiani, P. (2003). Bayesian methods. In M.B. Hand (Ed.), *Intelligent data analysis: An introduction* (pp. 128-166). New York: Springer.

Sebastiani, P., & Ramoni, M. (2001). Bayesian selection of decomposable models with incomplete data. *J Am Stat Assoc, 96*(456), 1375-1386.

Singh, M., & Valtorta, M. (1995). Construction of Bayesian network structures from data: A brief survey and an efficient algorithm. *Int J Approx Reason, 12*, 111-131.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. New York: John Wiley & Sons.

Wright, S. (1923). The theory of path coefficients: A reply to Niles' criticisms. *Genetics, 8*, 239-255.

Wright, S. (1934). The method of path coefficients. *Ann Math Statist, 5*, 161-215.

KEY TERMS

Bayes Factor: Ratio between the probability of the observed data under one hypothesis divided by its probability under an alternative hypothesis.

Conditional Independence: Let X, Y, and Z be three sets of random variables; then X and Y are said to be conditionally independent given Z, if and only if $p(x|z,y)=p(x|z)$ for all possible values x, y, and z of X, Y, and Z.

Directed Acyclic Graph (DAG): A graph with directed arcs containing no cycles; in this type of graph, for any node, there is no directed path returning to it.

Probabilistic Graphical Model: A graph with nodes representing stochastic variables annotated by probability distributions and representing assumptions of conditional independence among its variables.

Statistical Independence: Let X and Y be two disjoint sets of random variables; then X is said to be independent of Y, if and only if $p(x)=p(x|y)$ for all possible values x and y of X and Y.

This work was previously published in Encyclopedia of Data Warehousing and Mining, edited by J. Wang, pp. 674-677, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).