Probability Assessment with Maximum Entropy in Bayesian Networks

Wim Wiegerinck **Tom Heskes**
SNN, University of Nijmegen,
Geert Grooteplein 21, 6525 EZ, Nijmegen, The Netherlands
wimw@mbfys.kun.nl

Abstract

Bayesian networks are widely accepted as tools for probabilistic modeling in the medical domain. In modeling Bayesian networks in collaboration with domain experts, the definition of the network structure is relatively easy. The assessment of the conditional probability tables (CPT) is often a much more difficult task, even though there is a lot of statistical information available in the medical literature. The problem is twofold. In the first place it is usually not possible to use this information directly to fill in the CPTs. In the second place, the information is usually insufficient for a unique definition of the CPTs. A standard approach to define a probabilistic model on the basis of insufficient statistical information is to apply the Maximum Entropy Method (MaxEnt). MaxEnt searches for the unique model that maximizes the entropy under the constraints that it satisfies the given statistical information. In standard applications of MaxEnt for models defined by one joint probability table, these constraints are linear in the table entries. However, if MaxEnt is applied to a Bayesian network, i.e. the joint distribution is factorized into a product of CPTs, these constraints are typically nonlinear in the CPTs. In this paper we show how these nonlinear constraints can be dealt with, and we describe an algorithm that (locally) maximizes entropy under constraints in Bayesian networks. The method is illustrated by an example.

1 Introduction

Computer-based diagnostic decision support systems will play an increasingly important role in health care. They may improve the quality of the diagnostic process in accuracy and efficiency, while costs and burden of patients may be reduced. In addition, they can play an invaluable role in medical education. Potential users include general internists, super specialists, residents in internal medicine, and medical students.

The modern view is that decision support systems should be based on a probabilistic model. This approach has the advantage that it can deal with uncertainty in a consistent and mathematically correct way. In particular Bayesian networks[5; 3] provide a powerful and conceptually transparent formalism for probabilistic modeling.

Modeling of a Bayesian network consists of two parts, a qualitative and a quantitative part. The qualitative part is the determination of the structure of the network. If the network is build in collaboration with domain experts, the determination of the structure is often considered as a relatively easy task, since this task usually fits well with knowledge that medical experts often have about causal relationships between variables. The quantitative part consists of quantifying the conditional probability tables (CPTs) in the network. This part is often considered by medical experts as a much harder or even impossible task [2]. The reason is that medical domain experts themselves often have no idea about these probabilities. In most medical domains some statistical information $\mathcal{I}$ is provided in the literature. In such a case, one may try to choose the CPTs in the network such that network fits with $\mathcal{I}$. Unfortunately, $\mathcal{I}$ often does not translate directly into network CPTs, that is to say, it is often not clear to the experts how $\mathcal{I}$ should be translated into quantitative CPTs in the Bayesian network. Typically, $\mathcal{I}$ consists of conditional probabilities in the 'wrong direction', from 'effect' to 'cause'. In addition, these 'reversed' CPTs are often insufficient to uniquely define the desired CPTs in the network. The toy problem in the last section in the paper is an example where $\mathcal{I}$ has wrong direction and is insufficient for unique determination of the model. Often $\mathcal{I}$ can be formulated as linear probabilistic constraints, i.e., constraints of the form $\sum_{\{x\}} p(x) f_\alpha(x) = 0$, and/or $\sum_{\{x\}} p(x) g_\beta(x) \leq 0$, where $p(x)$ is the (joint) probability distribution and $f_\alpha(x)$ and $g_\beta(x)$ are functions of the state space $\{x\} = \{x_1, \ldots, x_n\}$. A typical example is a constraint on the conditional probability $p(x_1 = a | x_2 = b) = c$ which can be expressed as

$\sum_x p(x)(\delta_{x_1 a}\delta_{x_2 b} - c\delta_{x_2 b}) = 0$, where we used the Kronecker delta ($\delta_{xy} = 1$ if $x = y$ and $\delta_{xy} = 0$ if $x \neq y$).

In this paper, it is assumed that $\mathcal{I}$ is consistent, i.e. that there is at least one parameter setting of the distribution that satisfies the constraints. However, since $\mathcal{I}$ is in general insufficient for a unique determination of the model p, a whole set of distributions will satisfy the constraints. A standard way to proceed is to select a representative of this set of distributions by applying the Maximum Entropy Method (MaxEnt) [4]. MaxEnt searches the distribution that maximizes entropy under the given constraints. Roughly spoken, it selects the distribution p that satisfies the constraints without introducing any additional information.

In this paper, we apply MaxEnt to a Bayesian network with a given structure $p(x) = \prod_i p(x_i|\pi_i)$ to quantify its CPTs. The difference with MaxEnt applied to a general model $p(x)$ is that MaxEnt applied to a Bayesian network has to deal with a set of constraints *and* a set of independency statements. One approach could be to try to formulate the independency statements as as additional constraints to a general model $p(x)$ and apply standard MaxEnt to p. The way we proceed is, however, to keep the factorization into CPTs, and try to find the CPTs that maximizes the entropy of the joint distribution. As a consequence, a technical difference with standard MaxEnt is that the constraints, which are linear in the joint probability $p(x)$, are *non-linear* in the CPTs $p(x_i|\pi_i)$. This causes some complications in the optimization scheme of MaxEnt. However, one can effectively deal with these complications.

This workshop paper is organized as follows. In section 2 standard MaxEnt is shortly reviewed. In section 3, we show how the method applies to Bayesian networks. In section 4, the method is applied to a toy problem. We end the paper with a short discussion in section 5.

2 Maximum Entropy (MaxEnt)

In this section, we shortly review the standard Maximum Entropy (MaxEnt) method with linear probabilistic constraints [4]. We consider probability distributions $p(x)$ on a set of discrete variables $x = x_1, \ldots, x_n$ with a finite domain, $x_i \in \{1, \ldots, n_i\}$. If a set of linear constraints on p,

$$\sum_{\{x\}} f_\alpha(x)p(x) = 0 \qquad \alpha = 1 \ldots k \tag{1}$$

$$\sum_{\{x\}} f_\alpha(x)p(x) \geq 0 \qquad \alpha = k+1 \ldots m \tag{2}$$

is given, MaxEnt tries to find the probability distribution $p(x)$ that maximizes the entropy

$$H(p) = -\sum_{\{x\}} p(x) \log p(x) \tag{3}$$

under these constraints.

Introducing Lagrange multipliers $\lambda = \{\lambda_\alpha\}$, and a Lagrange multiplier γ to ensure normalization of p, we can formulate the optimization problem by Lagrangian

$$L(p, \lambda, \gamma) = H(p) + \sum_\alpha \sum_{\{x\}} \lambda_\alpha f_\alpha(x)p(x) + \gamma(\sum_{\{x\}} p(x) - 1) \tag{4}$$

which should be maximized with respect to p and minimized with respect to the Lagrange multipliers λ (within the domain $\lambda_\alpha \leq 0$ for $\alpha > k$) and γ. Taking the gradient of L with respect to $p(x)$, setting it to zero, and eliminating γ, we can solve p as explicitely as a function of λ. The solution p^* has the well known exponential form

$$p^*(x) = \frac{1}{Z} \exp \sum_\alpha \lambda_\alpha f_\alpha(x) \tag{5}$$

where Z is a proper normalization constant resulting from elimination of γ. Now we substitute the solution p^* into the Lagrangian L, which now becomes a function of λ only,

$$F(\lambda) = H(p^*) + \sum_\alpha \sum_{\{x\}} \lambda_\alpha f_\alpha(x)p^*(x) \tag{6}$$

which has to be optimized numerically , leading to the solution λ^*. According to the theory of Lagrange multipliers, the constrained optimization problem is now solved by the distribution p^* at λ^*.

3 MaxEnt in Bayesian networks

In this section, we show how the MaxEnt method under linear probabilistic constraints operates for a Bayesian network

$$p(x) = \prod_i p(x_i|\pi_i) \tag{7}$$

Again we want the maximize the entropy

$$H(p) = -\sum_{\{x\}} p(x) \log p(x) \tag{8}$$

under a set of linear constraints in p

$$\sum_{\{x\}} f_\alpha(x_\alpha)p(x) = 0 \qquad \alpha = 1 \ldots k \tag{9}$$

$$\sum_{\{x\}} f_\alpha(x_\alpha)p(x) \geq 0 \qquad \alpha = k+1 \ldots m \tag{10}$$

In which $f_\alpha(x_\alpha)$ is a function that depends on a subset of variables x_α. Introducing Lagrange multipliers λ_α for these constraints, and γ_i for normalization of the CPTs we can can formulate the optimization problem with the Lagrangian

$$L(\{p_i\}, \lambda, \gamma) = H(p) + \sum_\alpha \sum_{\{x\}} \lambda_\alpha f_\alpha(x_\alpha)p(x) + \gamma_i(\pi_i)(\sum_{\{x\}} p_i(x_i|\pi_i) - 1) \tag{11}$$

which should be maximized with respect to the CPTs $\{p_i\}$ and minimized with respect to the Lagrange multipliers. Now we cannot solve $\{p_i\}$ directly by taking the gradient of (11) with respect to all parameters, since this would only lead to a set of coupled non-linear equations for the CPTs.

What we can do, however, is taking the gradient of L with respect to a single CPT $p_i(x_i|\pi_i)$, for fixed λ and remaining CPTs$\{p_j\}_{j\neq i}$. Setting the gradient to zero, and eliminating $\gamma_i(\pi_i)$, we again get an *explicit* solution of $p_i(x_i|\pi_i)$, as a function of λ and the remaining CPTs $\{p_j\}_{j\neq i}$:

$$p_i^*(x_i|\pi_i) = \frac{1}{Z_i(\pi_i)} \exp\left\langle \sum_{\alpha\in C_i} \lambda_\alpha f_\alpha(x_\alpha) - \sum_{j\in D_i} \log p_j(x_j|\pi_j) \right\rangle_{x_i,\pi_i} \quad (12)$$

The average $\langle \ldots \rangle_{x_i,\pi_i}$ is taken with respect to the conditional distribution $p(x|x_i,\pi_i)$ (which only depend on the CPTs $\{p_j\}_{j\neq i}$). In (12), C_i is the subset of the constraints α such that the distribution of x_α depends on the state of x_i. In other words, $\alpha \notin C_i$ implies $p(x_\alpha|x_i) = p(x_\alpha)$. In a similar way, D_i is the subset of the nodes $j \neq i$, such that the child-parent combinations $\{x_j, \pi_j\}$ depends on the state of x_i. Again, in other words, $j \notin D_i$ implies $p(x_j, \pi_j|x_i) = p(x_j, \pi_j)$. Finally, $Z_i(\pi_i)$ are normalization constants for the CPTs.

Since the solution (12) is unique, it corresponds to the global maximum of L given that the other CPTs (and the Lagrange multipliers) are fixed. This means that in a sequence where at each step different CPTs are selected and updated (while keeping λ fixed) , the Lagrangian increases at each step (or remain constant). Since the Lagrangian is bounded for fixed λ we conclude that this iteration over all clusters of CPTs leads to a local maximum of L.

To find saddle points of L we propose the following two-step gradient descent procedure.

Initialization

- Initialize with random λ and random CPTs $\{p_i\}$.
- Fix λ and iterate (12) sequentially until a local maximum of $\{p_i\}$ is obtained.

λ - step Fix the CPTs and take a λ step into the direction of the negative gradient of the Lagrangian (some components, related to inequality constraints, will be set equal to zero if this step would push them outside their domain).

p -step Fix λ and iterate (12) sequentially, with CPTs initialized at their previous values, until convergence is reached.

In this way, we minimize with respect to λ in its domain, while remaining on a ridge $\partial L/\partial p_i(x_i|\pi_i) = 0$. If we converge, we obtain a local maximum of the entropy under the required constraints.

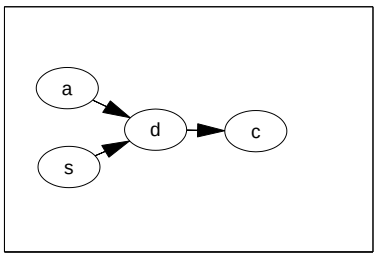

Figure 1: Structure of the network for coronary heart disease with four variables: *age* (a), *sex* (s), *heart-disease* (d), and *chest-pain* (c)

sex	age	asympt	non-AP	atyp-AP	typ-AP
m	30-39	1.9	5.2	21.8	67.7
m	40-49	5.5	14.1	46.1	87.3
m	50-59	9.7	21.5	58.9	92.0
m	60-69	12.3	28.1	67.1	94.3
f	30-39	0.3	0.8	4.2	25.8
f	40-49	1.0	2.8	13.3	55.2
f	50-59	3.2	8.4	32.4	79.4
f	60-69	7.5	18.6	54.4	90.6

Table 1: Conditional probabilities (percentages) of heart disease given age, sex and type of chest-pain (asymptomatic, non-AP pain, atypical AP-pain, typical AP-pain). This table is taken from literature and served as a constraint for the probability model in figure 1

4 An example: coronary heart disease

We illustrate the method by example involving the diagnosis of coronary heart disease, taken from [1]. In this example, we have four variables: *age* (a), *sex* (s), *heart-disease* (d), and *chest-pain* (c) . Following the example, *age* has four states (30-39, 40-49, 50-59, 60-69), *sex* has two states (*male*, *female*), *heart-disease* has two states (*true*, *false*), and *chest-pain* has four states (asymptomatic, non-AP pain, atypical AP-pain, typical AP-pain). We build a graphical structure according to figure 1.

The information that we have is a probability table $q(d|a,s,c)$ with conditional probabilities for all states of d, a, s, c, tabulated in table 1. Furthermore, there is no information, but we assume that we have the additional information that s and a are homogeneously distributed. The constraints $p(d|a,s,c) = q(d|a,s,c)$, $p(a) = 0.25$, $p(s) = 0.5$ are insufficient to uniquely specify the CPTs $p(d|a,s)$ and $p(c|d)$. We have applied MaxEnt to this problem. The CPTs that we obtained in this way are given in tables 2 and 3.

5 Conclusion and future work

If direct quantitative assessment of CPTs is to difficult for domain experts, and if other statistical information about the domain is available, but in the 'wrong

age	male	female
30-39	19	4
40-49	42	12
50-59	55	29
60-69	64	51

Table 2: Conditional probabilities (percentages) of heart disease (d = true) conditioned on age and sex. These CPTs are obtained by MaxEnt.

d	asympt.	non AP	atypical AP	typical AP
true	3	7	35	55
false	31	33	30	6

Table 3: Conditional probabilities (percentages) of having a certain state of chest pain (asymptomatic, non-AP pain, atypical AP-pain, typical AP-pain), given the state of heart disease (true or false). This CPTs is obtained by MaxEnt.

direction' and insufficient to uniquely define the desired CPTs in the network, then MaxEnt for Bayesian networks may provide a useful method for assessment of the quantitative CPTs. MaxEnt is not the only method for quantitative assessment of CPTs. Other methods have been proposed previously [2]. One of the features of MaxEnt for Bayesian networks is that the optimization procedure requires only local computations (if the constraints are local, i.e. involve only a few variables). This feature is crucial for application to large scale models.

Currently we collaborate with domain experts to study the feasibility of the construction of large scale Bayesian networks for medical diagnosis. Typically these networks will consists of several hunderds of nodes. One of the bottlenecks is the quantitative assessment of CPTs in these network, for reasons described in this paper. Our future work will include the study of the practical usefulness of the MaxEnt method for quantitative assessment of CPTs in such models.

Acknowledgements

This project is funded by the Dutch Technology Foundation STW. Jan Neijt is thanked for pointing us at the medical example.

References

[1] *Diagnostisch Kompas.* 1997.

[2] Marek J. Druzdzel and Linda C. van der Gaag. Elicitation of probabilities for belief networks: Combining qualitative and quantitative information. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 141–148, 1995.

[3] F.V. Jensen. *An introduction to Bayesian networks.* UCL Press, 1996.

[4] R. Levine and M. Tribus, editors. *The Maximum Entropy Formalism.* 1979.

[5] J. Pearl. *Probabilistic Reasoning in Intelligent systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers, Inc., 1988.