Variational Approximations in a Broad and Detailed Probabilistic Model for Medical Diagnosis

WAJJ Wiegerinck [1] EWMT ter Braak [2] WJPP ter Burg [3]
MJ Nijman [1] YL O [3] JP Neijt [2] and HJ Kappen [1]

(1) Stichting Neurale Netwerken, University of Nijmegen
Nijmegen, The Netherlands

(2) Dept. of Internal Medicine, Utrecht University Hospital
Utrecht, the Netherlands

(3) CI & A, Utrecht University Hospital
Utrecht, The Netherlands

Abstract

Exact inference in large, detailed probabilistic models for medical diagnosis is typically computationally infeasible, and approximate schemes are therefore of great importance. In this paper, we consider variational methods, which provide bounds on the probabilities of interest. We sketch some characteristics of a typical broad and detailed probabilistic model (BDPM) for medical diagnosis and describe how recently developed variational techniques can be applied for approximate inference in such a model. Currently we are developing a BDPM to study the practical feasibility and the usefulness of a system based on such a model in medical practice.

1 Introduction

Computer-based diagnostic systems can play many roles in decision support and other areas of medical practice. Most systems are designed to produce a differential diagnosis using a set of input findings entered by the user (as opposed to textbooks that tend to do the reverse - taking individual diseases and listing the associated findings). At present several systems are available such as Meditel [1], Quick Medical Reference (QMR) [2], DXplain [3], and Iliad [4].

The different systems that have been developed so-far use a variety of modelling approaches which can be roughly divided into two categories: rule-based approaches with or without uncertainty and probabilistic methods. The rule based approach can be viewed as an attempt to simplify the probabilistic approach in order to reduce computational complexity. The probabilistic approach, however, has the advantage of mathematical consistency and correctness. In particular belief networks (see e.g. [5, 6]) provide a powerful and conceptual transparent formalism for probabilistic modelling. The progress that

has been made during the last decade in exact computation in belief networks makes the argument in favour of rule based approaches less and less persuasive. Indeed, most modern approaches for medical diagnosis are based on the probabilistic approach.

The inadequacies of the current systems [7, 8] are therefore not due to the method used, but rather due to the scope and level of detail at which the disease areas are modelled. Either the system is based on detailed modelling of a restricted medical sub-domain [9, 10], or the system covers a large domain, at the expense of the level of detail at which the disease areas are modelled [11]. The reason for this restriction is that standard belief networks become intractable for exact computation if a large medical area would be modelled in detail.

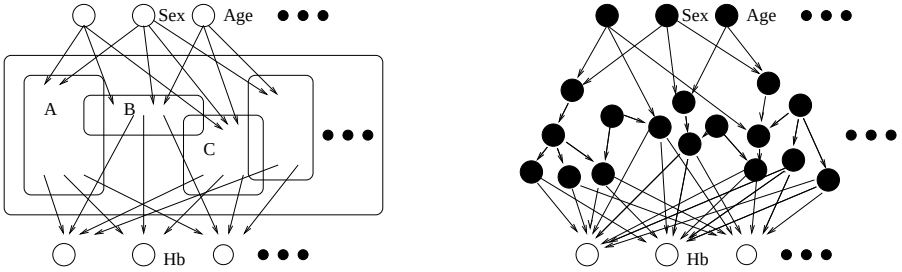
To proceed one has to rely on *approximate* computations. Recently, variational methods for approximation are becoming increasingly popular [12, 13, 14, 15, 16]. An advantage of variational methods techniques is that they provide guaranteed bounds on the level of approximations in contrast to stochastic sampling methods [15], which may yield unreliable results due to finite sampling times. Until now, variational approximations have been less widely applied than Monte Carlo methods, arguably since their use is not universally so straightforward. In this paper, however, we will argue that variational methods are indeed applicable to large, detailed belief networks for medical diagnosis constructed by human experts. In particular we will argue that such models will typically have a bipartite structure that is intractable for exact inference.

The paper is organised as follows. In section 2 we discuss the bipartite structure of a broad and detailed probabilistic model for medical diagnosis constructed by human experts. In section 3 we describe bipartite networks in mathematical terms. In section 4 we show how variational methods can provide lower and upper bounds on quantities of interest in bipartite networks. In section 5 we give a short description of a demo of a medical system based on the methods described in this paper. We conclude with a discussion in section 6.

2 Modelling and network structure

We here outline how the structure of a broad and detailed belief network constructed by human experts will typically look like, based on an extrapolation of the current modelling experiences of the physicians in our group. Details of the medical domain, the choice of variables, the inclusion of pathophysiological hidden variables, the transformation of continuous variables into binary variables, the assessment of probabilities etc. are beyond the scope of this paper and will be discussed elsewhere.

Medical experts tend to subdivide knowledge concerning a medical domain, e.g. anaemia, into sub-domains with a relatively small overlap. As a result, the network that models the full domain will typically have a modular structure (cf. fig. 1). Each module represents knowledge about a sub-domain and is modelled by a reasonably small belief network in which the nodes have only a small number of parents. The interconnectivity between the sub-domains is also small. There are two types of variables outside the sub-domains which link the many sub-domains together. One such type are variables like 'age' or 'sex', which determine prior probabilities of diseases. These variables are common ancestors of a large number of sub-domains. The other type of variables are influenced by causes in many sub-domains. An example is the variable 'hemoglobin level' (Hb) in the domain of anaemia. There are many sub-domains within anaemia, each impacting on Hb. That is, variables like Hb are common children of a large number of sub-domains. Since



(a) Modular structure (b) Graphical structure

Figure 1: Modular and graphical network structure. Left: modular structure of the network. A, B, C ... represent (overlapping) sub-domains. Each sub-domain is modelled by a number of nodes (cf. right figure) representing variables that are relevant in that domain. The upper nodes, e.g. 'sex' and 'age' represent common ancestors of nodes in several sub-domains. The lower nodes, e.g. 'Hb' represent common children of nodes in several sub-domains (e.g. related to anaemia). Right: underlying bipartite graphical structure of same network. Filled circles: nodes in sub-domains and their common ancestors. Open circles: common children

these nodes then have parents in many sub-domains, modelling using explicit probability tables is not feasible (the sizes of the tables grow exponentially in the number of parents). Fortunately, it is not necessary to define these large tables explicitly, since medical experts are likely to have in mind a more compact functional relation between these variables and their direct parents. Such a compact relationship is typically a noisy-OR [5] or a similar parametrised relationship.

2.1 Inference

The inference problem is to compute probabilities in the model, given evidence. If a network includes only a few medical sub-domains, exact inference using standard algorithms is feasible. However, in a detailed *and* broad network, such as described in the previous section, exact inference is infeasible due to the connectivity between the modules via the shared variables such as Hb. For instance, exact inference involves a summation over all the parent-states of the variable Hb and this exponentially large summation cannot be performed efficiently.

3 Biparite networks

The considerations of the previous section motivates us to consider a bipartite [1] network $\mathcal{N}$. A bipartite network consists of two parts (see fig. 1(b)). One part, which we call the parent network $\mathcal{N}_{\text{parent}}$, is a conventional belief network (black nodes). The only restriction for this parent network is that it is assumed to be tractable. Nodes of the parent network are connected (via noisy-OR gates [5]) to the common children in a second network, which we call the child network $\mathcal{N}_{\text{child}}$ (white nodes in fig. 1(b)). Conditioned on the state of the parent network, the nodes in $\mathcal{N}_{\text{child}}$ are independent. One may view a bipartite network as a generalisation of the bipartite network of the QMR-DT database [11, 13, 17]. The

[1] Note that our definition of bipartite networks differs from the usual definition in graph theory

difference is that the upper network in QMR-DT consists merely of disconnected nodes, while in our case the upper layer is a belief network with a non-trivial graphical structure.

In a bipartite network the probability of a state $S = (S_1, \ldots, S_n)$ therefore assumes the following factorised form,

$$P(S) = \prod_k P(S_k|S_{\pi_k}) \prod_i nOR(S_i|S_{\pi_i}) \tag{1}$$

with $i \in \mathcal{N}_{\text{child}}$ and $k \in \mathcal{N}_{\text{parent}}$. We denote S_{π_k} for the state of the parent set π_k of the node k. All the parents are in $\mathcal{N}_{\text{parent}}$. For convenience we focus in this treatment on binary variables $S_i \in \{0, 1\}$, representing the presence or absence of a disease or a finding. We will use shorthand notation S_i^- for $S_i = 0$, S_i^+ for $S_i = 1$. The noisy-OR gates $nOR(S_i|S_{\pi_i})$ are defined such that the probability for S_i^- is

$$nOR(S_i^-|S_{\pi_i}) \;=\; (1 - q_{i0}) \prod_{j \in \pi_i} (1 - q_{ij})^{S_j} \tag{2}$$

and $nOR(S_i^+|S_{\pi_i}) = 1 - nOR(S_i^-|S_{\pi_i})$. The parameter q_{i0} is the so-called leak probability. This is the probability of a positive finding S_i^+ if all its parents are 0. The parameter q_{ij} can be interpreted as the probability on S_i^+ if only the parent $S_j = 1$, while the others are 0 (if there was no leak probability).

Since $\mathcal{N}_{\text{parent}}$ is assumed to be tractable, exact computation is efficient if the nodes in $\mathcal{N}_{\text{child}}$ are not observed. Also negative findings S_i^- can be dealt with in linear time, since $nOR(S_i^-|S_{\pi_i})$ factorises over the parents (see (1)). The problem in this network is that exact computation is inefficient for positive nodes S_i^+ in the child network[13, 17], since the computational costs of inference involving these nodes scales exponentially in the number of parent states. In the next section we will propose variational approximations to deal with this problem.

4 Variational methods for bipartite networks

The starting point of variational methods is to transform the inference problem into an equivalent optimisation problem. The optimisation problem has a simpler structure than the original inference problem, but there are unknown parameters involved. If one is able to find the optimal set of these parameters, one has solved the inference problem. Of course, to find the optimal set of parameter is just as hard, or even harder than the inference problem itself. The trick in the variational methods is to restrict the parameter space so that the optimisation problem becomes feasible. Although the solution of the restricted optimisation problem does not lead to the exact solution of the inference problem, it is guaranteed to bound the exact solution.

In the following subsections, we show that upper and lower bounds of marginal likelihoods can be computed. The way that these bounds are derived are similar to the upper bounds for the bipartite QMR-DT network derived in [13, 17] and the lower bounds for sigmoid belief networks derived in [12]. The difference with these previous papers, however, is that we are able to exploit more fully the graphical structure of the parent network, as in [18, 19], leading to a better approximation algorithm.

Once upper and lower bounds of marginals are computed, bounds on conditional probabilities can be obtained by taking fractions of the marginal bounds, see [13, 17] for more details. Techniques to combine approximate and exact combinations can also be found in these references.

4.1 Upper bound

Before we proceed, it is convenient to re-express the noisy-OR gates (2) using an exponential notation with parameters $\theta_{ij} = -\log(1 - q_{ij})$,

$$\begin{aligned} nOR(S_i^-|S_{\pi_i}) &= \exp(-\theta_{i0} - \sum_{j\in\pi_i} \theta_{ij} S_j) \\ &= \exp(-z_i) \end{aligned} \tag{3}$$

in which with $z_i = \sum_j \theta_{ij} S_j + \theta_{i0}$. The upper bound of the marginal likelihood is based on the following inequality,

$$\ln(1 - e^{-z}) \leq \xi z - F^*(\xi) \tag{4}$$

in which $F^*(\xi) = -\xi \ln \xi + (1 + \xi)\ln(1 + \xi)$. For given z, this inequality is valid for each value of the variational parameter ξ. If the right-hand-side of (4) is minimised with respect to ξ, the inequality becomes an equality (for given z).

Applying inequality (4) to the bipartite network (1) we obtain the following bound on the marginal likelihood $P(S_V)$ of the 'visible' variables $V = \{V_{\text{parent}}, V_{\text{child}}\} \subset \{\mathcal{N}_{\text{parent}}, \mathcal{N}_{\text{child}}\}$,

$$P(S_V) \leq \sum_{\{S_{H\text{parent}}\}} \prod_k P(S_k|S_{\pi_k}) \exp(\sum_{i^+} \xi_{i^+} z_{i^+} - F^*(\xi_{i^+}) - \sum_{i^-} z_{i^-}) \tag{5}$$

As before, nodes with indices k are in $\mathcal{N}_{\text{parent}}$. Nodes with indices $i^{+/-} \in V_{\text{child}}$ have positive/negative findings. $S_{H\text{parent}}$ are undetermined (hidden) states in the parent network. Note that this bound is tractable, since the product over child nodes is log-linear in the parent states S_j. The graphical structure of the parent states is therefore not affected, and remains tractable. To get the bound as tight as possible, we optimise the right-hand-side of (5) with respect to the ξ_{i^+}'s, using some numerical procedure.

4.2 Lower bound

Recently it has been proposed to use variational techniques to obtain a lower bound of the marginal likelihood in sigmoid belief networks [12, 18, 19]. Similar methods can be applied to the bipartite networks considered in this paper. To derive the lower bound for a network with noisy-OR gates, we use the following expansion of the exponential function [17]

$$1 - \exp(-z) = \prod_{\kappa=0}^{\infty} (1 + \exp(-2^\kappa z))^{-1}$$

from which we deduce, usings Jensen's inequality,

$$\langle \ln(1 - \exp(-z)) \rangle_Q \geq -\sum_{\kappa=0}^{\infty} \ln(\langle 1 + \exp(-2^\kappa z) \rangle_Q).$$

Using this bound, combined with the general theory of variational lower bounds of the likelihood we obtian,

$$\begin{aligned} \ln P(S_V) \geq \mathcal{F}_V[Q] = &\sum_k \langle \ln P(S_k|S_{\pi_k} \rangle_Q - \sum_{i^-} \langle z_{i^-} \rangle_Q \\ &+ \sum_{i^+} \sum_{\kappa=0}^{\infty} \ln(1 + \langle \exp(-2^\kappa z_{i^+}) \rangle_Q) \\ &- \sum_{\{S_H\}} Q(S_H) \ln Q(S_H) \end{aligned} \tag{6}$$

which is valid for any approximating distribution $Q(S_H)$. As before, $k \in \mathcal{N}_{\text{parent}}$ and $i \in \mathcal{N}_{\text{child}}$, $z_i = \sum_j J_{ij} S_j + h_i$ and $\langle \cdot \rangle_Q$ is the average with respect to the so-called mean field distribution $Q(S_H)$. S_H are again the undetermined states in the parent network. Since this inequality holds for any Q, one can make the bound as tight as possible by optimising $\mathcal{F}_V[Q]$ with respect to Q. Recently, it has been shown that if tractable belief networks are used for the mean field distribution $Q(S_H)$, then the optimisation of can be performed efficiently using mean field equations [18, 19]. The same studies noted the increase in precision of the bound if the structure of the mean field distribution had more overlap with the structure of the network that was to be approximated. In the the bipartite networks considered in this paper, a natural structure for the mean field distribution $Q(S_H)$ is a belief network with the same structure as $\mathcal{N}_{\text{parent}}$.

5 Demo

In this paper, we have laid the theoretical foundations of a BDPM based on a bipartite network. In order to examine its usefulness to support medical practice, we are currently developing such a system for the domain of anaemia. In the near future, the system will be evaluated in a clinically realistic setting. Performance will be assessed not only in terms of diagnostic accuracy, but more importantly how the system influences the performance of physicians using the system, and to what extent the users are satisfied with the system [7].

The BDPM engine is connected to a user friendly interface. In addition to a differential diagnosis based on the available patient information, the interface will also provide information, such as abstracts and references to relevant literature, explanatory functions to motivate the conclusions of the system, and advice for further action. It is our aim to use this demo both in a clinical and an educational setting.

6 Discussion

The development of an automated system for comprehensive medical diagnosis in internal medicine represents a great challenge for AI. A broad and detailed probabilistic network is intractable for exact inference in this context, although recent developments in variational methods may provide a practical solution for approximate inference. Building on the theoretical foundations established here, it is a topic of vital importance to assess their performance in a real world situation.

Acknowledgements

This research is supported by the Technology Foundation STW, applied science division of NWO and the technology programme of the Ministry of Economic Affairs

References

[1] Meditel, Devon, Pa. *Meditel: Computer assisted diagnosis*, 1991.

[2] CAMDAT, Pittsburgh. *QMR (Quick Medical Reference)*, 1992.

[3] Massachussetts General Hospital, Boston. *DXPLAIN*, 1992.

[4] Applied Informatics, Salt Lake City. *ILIAD*, 1992.

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

[6] E. Castillo, J. M. Gutierrez, and A. S. Hadi. *Expert Systems and Probabilistic Network Models*. Springer, 1997.

[7] AS Elstein, CP Friedman, FM Wolf, G Murphy, J Miller, P Fine, P Heckerling, T Miller, J Sisson, S Barlas, K Biolsi, M Ng, X Mei, and T Franz. Performance of four computer-based diagnostic systems. *N-Engl-J-Med.*, 330(25):1792–6, 1994.

[8] ES Berner, JR Jackson, and J Algina. Relationships among performance scores of four diagnostic decision support systems. *J-Am-Med-Inform-Assoc.*, 3(3):208–15, 1996.

[9] D.E. Heckerman, E.J. Horvitz, and B.N. Nathwani. Towards normative expert systems: part I, the Pathfinder project. *Methods of Information in medicine*, 31:90–105, 1992.

[10] D.E. Heckerman and B.N. Nathwani. Towards normative expert systems: part II, probability-based representations for efficient knowledge acquisition and inference. *Methods of Information in medicine*, 31:106–116, 1992.

[11] M.A Shwe, B. Middleton, D.E. Heckerman, M. Henrion, Horvitz E.J., H.P. Lehman, and G.F. Cooper. Probabilistic Diagnosis Using a Reformulation of the Internist-1/ QMR Knowledge Base. *Methods of Information in Medicine*, 30:241–55, 1991.

[12] L.K. Saul, T. Jaakkola, and M.I. Jordan. Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4:61–76, 1996.

[13] T.S. Jaakkola and M.I. Jordan. Variational methods and the QMR-DT database. MIT Computational Cognitive Science Technical Report 9701, Massachusetts Institute of Technology, 1997.

[14] T. S. Jaakkola and M. I. Jordan. Recursive Algorithms for Approximating Probabilities in Graphical Models. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996.

[15] M. I. Jordan, editor. *Learning in Graphical Models*, volume 89 of *NATO ASI, Series D: Behavioural and Social Sciences*. Kluwer, 1998.

[16] H.J. Kappen and F.B. Rodríguez. Boltzmann machine learning using mean field theory and linear response correction. In Micheal Kearns, editor, *Advances in Neural Information Processing Systems*, volume 10. The MIT Press, 1998.

[17] T. Jaakkola. *Variational Methods for Inference and Estimation in Graphical Models*. PhD thesis, Massachusetts Institute of Technology, 1997.

[18] W. Wiegerinck and D. Barber. Mean field theory based on belief networks for approximate inference. In *ICANN'98: International Conference on Artificial Neural Networks, Skövde*, 1998.

[19] W. Wiegerinck and D. Barber. Variational belief networks for approximate inference. In *Tenth Netherlands/Belgium Conference on Artificial Intelligence (NAIC'98)*. CWI, 1998. in press.