Inference in the Promedas medical expert system

Bastian Wemmenhove[1], Joris M. Mooij[1], Wim Wiegerinck[1], Martijn Leisink[1], Hilbert J. Kappen[1], and Jan P. Neijt[2]

[1] Department of Biophysics, Radboud University Nijmegen, 6525 EZ Nijmegen, The Netherlands

[2] Internal Medicine, University Hospital Utrecht Utrecht, the Netherlands

Abstract. In the current paper, the Promedas model for internal medicine, developed by our team, is introduced. The model is based on up-to-date medical knowledge and consists of approximately 2000 diagnoses, 1000 findings and 8600 connections between diagnoses and findings, covering a large part of internal medicine. We show that Belief Propagation (BP) can be successfully applied as approximate inference algorithm in the Promedas network. In some cases, however, we find errors that are too large for this application. We apply a recently developed method that improves the BP results by means of a loop expansion scheme. This method, termed Loop Corrected (LC) BP, is able to improve the marginal probabilities significantly, leaving a remaining error which is acceptable for the purpose of medical diagnosis.

1 Introduction

In this paper we present the Promedas medical diagnosis model. It is an expert system for doctors based on a Bayesian network structure for which the calculation of marginal probabilities is tractable for many cases encountered in practice. For those cases that are intractable (i.e. a junction tree algorithm is not applicable), alternative algorithms are required. A suitable candidate for this task is Belief Propagation (BP), which is a state-of-the art approximation method to efficiently compute marginal probabilities in large probability models [1, 2]. Over the last years, BP has been shown to outperform other methods in rather diverse and competitive application areas, such as error correcting codes [3, 4], low level vision [5], combinatoric optimization [6] and stereo vision [7].

In medical expert systems, so far the success of BP has been limited. Jaakkola and Jordan [8] successfully applied variational methods to the QMR-DT network [9] but BP was shown not to converge on these same problems [2]. We find that BP does converge on all Promedas cases studied in the current paper. Although this does not guarantee convergence in all possible cases, we note that double loop type extensions to BP [10] may be applied when convergence ceases. Here we compute the marginal errors of BP and apply a novel algorithm, termed Loop Corrected Belief Propagation (LCBP) [11] to cases in which the error becomes

unacceptable. We argue that this method potentially reduces the error to values acceptable for medical purposes.

Recently a company was founded that uses the Promedas network to develop a commercially available software package for medical diagnostic advise. A demonstration version can be downloaded from the website `www.promedas.nl`. The software will become available as a module in third party software such as laboratory or hospital information systems or stand alone designed to work in a hospital network to assist medical specialists. In all cases the software will be connected to some internally used patient information system. This year the Promedas software will be available via a web portal as well. This might be operational at the time of the AIME congress. Physicians can visit the website, enter medical characteristics of a specific case and immediately obtain a list of most probable diagnoses. The Promedas web portal uses the full available database of diagnoses and findings.

2 Inference in the Promedas graphical model

The global architecture of the diagnostic model in Promedas is similar to QMR-DT [9]. It consists of a diagnosis-layer that is connected to a layer with findings. Diagnoses (diseases) are modeled as a priori independent binary variables $d_j \in \{0,1\}$, $j \in \{1,\ldots,N_{\mathrm{D}}\}$, causing a set of symptoms or findings $f_i \in \{0,1\}$. In the user interface, a significant part of the findings are presented as continuous variables. These are discretized in a medically sensible way. The interaction between diagnoses and findings is modeled with a noisy-OR structure, indicating that each parent j has an individual probability of causing a certain finding i to be true if it is in the parent set $V(i)$ of i, and there is an independent probability λ_i that the finding is true without being caused by a parent (disease). Thus

$$\begin{aligned} p(f_i = 0|\mathbf{d}) &= [1-\lambda_i] \prod_{j \in V(i)} [1 - w_{ij} d_j] \\ p(f_i = 1|\mathbf{d}) &= 1 - p(f_i = 0|\mathbf{d}) \end{aligned} \tag{1}$$

The parameters $\{\lambda_i\}, \{w_{ij}\}$, together with the disease prevalences (ranging from 0.001 to 0.1) are the model parameters determined by the medical experts. The disease nodes are coupled to risk factors, such as, e.g., concurrent diagnoses and nutrition. Risk factors are assumed to be observed and to modify the prevalences of the diagnoses. From a database of model parameters, the graphical model and a user-interface for Promedas are automatically compiled. This automatic procedure greatly facilitates changes in the model, such as adding or removing diseases, as required in the design phase of the model. Once the graphical model has thus been generated, we use Bayesian inference to compute the probability of all diagnoses in the model given the patient data. Before computation, we remove all unclamped (i.e. unobserved) findings from the graph, and we absorb negative findings in the prevalences [8]. Only a network of positively clamped findings and their parents remain. Using standard techniques for the calculation of posterior

distributions directly on the factor graph in the above representation, either with a junction tree algorithm ([12]) or approximation techniques, is limited to cases in which the size $|V(i)|$ of the interaction factors is not too large. In Promedas, however, sets containing 30 nodes (i.e. findings that may have 30 different causes) are not uncommon. Thus it is helpful to reduce the maximum number of members of factor potentials, which may be achieved by adding extra (dummy) nodes to the graph [13, 14]. The version of Promedas that is studied in this paper contains over 10000 variables, including about 2000 diagnoses, and 8600 connections between diagnoses and findings.

Despite these measures computation can still be intractable when the number of positive patient findings becomes large [8]. In that case, we must resort to approximations. The feasibility of this approach is studied in the remainder of the paper. In the next section we first report results of applying Belief Propagation to a number of "virtual patient" cases, followed by tests of a version of the recently developed Loop Corrected Belief Propagation [15, 11] algorithm on these cases. The idea of LCBP can be understood as follows. BP is a method that is exact on graphs that are tree-like. This means that if one removes a node from the graph, the probability distribution on its neighbors (the so-called cavity distribution) factorizes. When BP is applied to graphs with loops, this is no longer true and the cavity distribution contains correlations. The LCBP method incorporates estimates of these correlations in a message passing scheme. For more details see [11].

3 Simulations with virtual patient data

Using the model first as a generator of virtual patients, we generated, 1000 patient cases with $N_{\mathrm{d}} = 1$ and another 1000 with $N_{\mathrm{d}} = 4$ where N_{d} represents the number of randomly selected true diseases for the generation of patient data. The first result we report is the fact that on all cases that we generated BP converged. This contrasts with previous results by Murphy et. al. [2], found for the QMR-DT network, where the small prior probabilities seemed to prevent convergence in a couple of complex cases. The maximal marginal errors in the BP results are typically small but may occasionally be rather large. In fig. 1 left we plot the error versus the tree width of the JT method, which is an indication of the complexity of the inference task. From fig. 1 left we conclude that the quality of the BP approximation is only mildly dependent on this complexity. It follows that for patient cases where exact computation is infeasible, BP gives a reliable alternative for most cases. To avoid cases where the error is unacceptably large we propose to use the so-called Loop Corrected BP method.

The right picture of fig. 1, displays results of applying LCBP to a set of 150 $N_d = 1$ virtual patient cases. Horizontally, the maximal error in BP single node marginals is plotted, and vertically the maximal error after applying the loop correction scheme. Only cases with nonzero error (i.e. loopy graphs) are plotted, 86 in total. The maximal error in the marginals produced by BP typically reduces

one order of magnitude after applying LCBP. The largest maximum error over all cases in this sample reduced from 0.275 to 0.023.

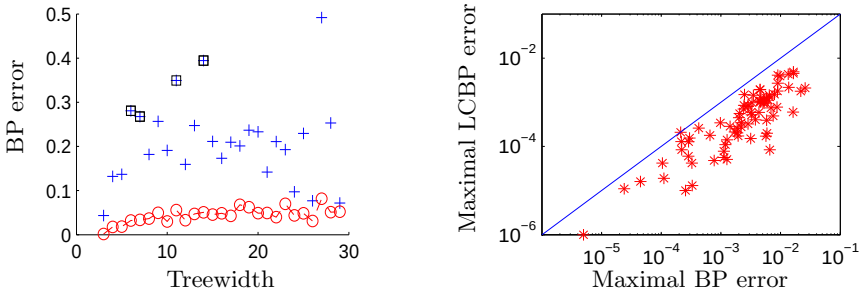

Fig. 1. *Left*: BP maximal error(∘) averaged over instances, largest maximal error (+) as a function of treewidth for $N_d = 4$. The squares mark instances with large error which we have later subjected to LCBP (see table 1). *Right*: $N_d = 1$ Maximal single node marginal error of LCBP (vertical) versus BP (horizontal). All data lie on the side of the line where the LCBP error is smaller than the BP error.

As a second test, we applied the method to a few cases in the left picture of figure 1, where we attempted to reduce large BP errors of these complex multiple disease errors. A drawback of the current implementation of LCBP is its rather large computation time when Markov blanket sizes grow large. For the implementation of LCBP that we used, computation time grows as N^2 (assuming constant maximal degree per node), but also grows exponentially in the number of nodes in the largest Markov blanket. The exponential scaling of the algorithm in Markov blanket size forced us to look at a few relatively easy cases only. Results for the BP errors marked by a black square in figure 1 are reported in table 1:

Table 1. LCBP results on complex instances with large errors:

Treewidth	rms error BP	max error BP	rms error LCBP	max error LCBP
6	0.0336	0.2806	0.0021	0.0197
7	0.0429	0.2677	0.0017	0.0102
11	0.0297	0.3494	intractable	intractable
14	0.0304	0.3944	0.0011	0.0139

The maximal error of LCBP clearly reduces to acceptable levels, but the computation time is prohibitive for complex cases. The solution to this problem may be an alternative implementation, taking into account only nontrivial correlations between pairs of variables in the Markov blanket (see [15]), and consequently scales polynomially in the Markov blanket size. We did not consider

this algorithm in the current investigation, since its implementation is much more involved, but the promising results obtained here motivate us to do so in the future.

4 Conclusions

In this paper we have shown that BP is an attractive alternative for exact inference for complex medical diagnosis inference tasks. In some isolated instances, BP produces large errors and we have shown that loop corrected BP can significantly reduce these errors. Therefore, for practical purposes it seems worthwile to further develop an efficient version of LCBP that scales polynomially in the Markov-blanket size, such as the one proposed in [15].

References

1. J. Pearl. *Probabilistic reasoning in intelligent systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, California, 1988.
2. Kevin P. Murphy, Yair Weiss, and Michael I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of Uncertainty in AI*, pages 467–475, 1999.
3. R.G. Gallager. *Low-density parity check codes*. MIT Press, 1963.
4. R. McElice, D. MacKay, and J. Cheng. Turbo decoding as an instance of pearl's belief propagation algorithm. *Journal of Selected Areas of Communication*, 16:140–152, 1998.
5. W.T. Freeman, E.C. Pasztor, and O.T. Carmichael. Learning low-level vision. *Int. J. Comp. Vision*, 40:25–47, 2000.
6. M. Mezard, G. Parisi, and R. Zecchina. Analytic and algorithmic solution of random satisfiability problems. *Science*, 297, 2002.
7. J. Sun, Y. Li, S.B. Kang, and H.-Y. Shum. Symmetric stereo matching for occlusion handling. *Proceedings CVPR*, 2:399–406, 2005.
8. T. Jaakkola and M.I. Jordan. Variational probabilistic inference and the QMR-DT network. *Journal of artificial intelligence research*, 10:291–322, 1999.
9. 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.
10. T. Heskes, K. Albers, and H.J. Kappen. Approximate inference and constraint optimisation. In *Proceedings UAI*, pages 313–320, 2003.
11. J.M. Mooij, B. Wemmenhove, H.J. Kappen, and T. Rizzo. Loop corrected belief propagation. In *Proceedings of AISTATS*, 2007.
12. S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilties on graphical structures and their application to expert systems. *J. Royal Statistical society B*, 50:154–227, 1988.
13. D. Heckerman. A tractable inference algorithm for diagnosing multiple diseases. In *Proceedings UAI*, pages 163–171. Elsevier Science, 1989.
14. M. Takinawa and B. D'Ambrosio. Multiplicative factorization of noisy-MAX. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence UAI99*, pages 622–30, 1999.
15. A. Montanari and T. Rizzo. How to compute loop corrections to the bethe approximation. *Journal of Statistical Mechanics*, page P10011, 2005.