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 large parts of internal medicine. Promedas is currently being evaluated informally by specialists in internal medicine at the Utrecht university hospital and is receiving positive responses. We show that Belief Propagation (BP) can be successfully applied to approximate inference problems in the Promedas network. BP converges on all patient test cases, which were generated with the help of the model itself. 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

Modern-day medical diagnosis is a very complex process, requiring accurate patient data, a profound understanding of the medical literature and many years of clinical experience. This situation applies particularly to internal medicine, because it covers an enormous range of diagnostic categories. As a result, internal medicine is differentiated in super-specializations.

Diagnosis is a process, by which a doctor searches for the cause (disease) that best explains the symptoms of a patient. The search process is sequential, in the sense that patient symptoms suggest some initial tests to be performed. Based on the outcome of these tests, a tentative hypothesis is formulated about the possible cause(s). Based on this hypothesis, subsequent tests are ordered to confirm or reject this hypothesis. The process may proceed in several iterations until the patient is finally diagnosed with sufficient certainty and the cause of the symptoms is established.

A significant part of the diagnostic process is standardized in the form of protocols. These are sets of rules that prescribe which tests to perform and in which order, based on the patient symptoms and previous test results. These

rules form a decision tree, whose nodes are intermediate stages in the diagnostic process and whose branches point to additional testing, depending on the current test results. The protocols are defined in each country by a committee of medical experts.

In the majority of the diagnoses that are encountered, the guidelines are sufficiently accurate to make the correct diagnosis. For these "routine" cases, a decision support system is not needed. In 10-20 % of the cases, however, the diagnostic process is more difficult. As a result of the uncertainty about the correct diagnosis and about the next actions to perform, the decisions made by different physicians at different stages of the diagnostic process do not always agree and lack "rationalization". In these cases, normally a particularly specialized colleague or the literature is consulted.

For these difficult cases computer based decision support may be of added value. Its strength is that it can give valuable information support for those patients of medical specialists that suffer from a disease that is outside his or her super-specialization. It may thus result in an improved and more rationalized diagnostic process, as well as higher efficiency and cost-effectiveness. The benefits of a successful decision support system for internal medicine could be far-reaching. Since 1996, we have been developing a clinical diagnostic decision support system for internal medicine, called Promedas. In this system, patient information, such as age and gender, and findings, such as symptoms, results from physical examination and laboratory tests can be entered. The system then generates patient-specific diagnostic advice in the form of a list of likely diagnoses and suggestions for additional laboratory tests that are expected to be particularly informative to establish or rule out any of the diagnoses considered.

The system is intended to support diagnostics in the setting of the outpatient clinic and for educational purposes. Its target users are general internists, super specialists (i.e. endocrinologists, rheumatologists, etc.), interns and residents, medical students and others working in the hospital environment.

The Promedas model, which we present in this paper, is based on a Bayesian network structure for which the calculation of marginal probabilities is tractable for almost all 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.

Promedas is still under development. A preliminary version of Promedas is currently being evaluated informally, during the weekly meeting of specialists in internal medicine at the Utrecht university hospital and is receiving positive responses.

2 Promedas, the model

The global architecture of 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. Findings are modeled as binary noisy-OR gates with relevant diagnoses as parents. In the user interface, a significant part of the findings are presented as continuous variables. These are discretized in a medically sensible way, and internally modeled as groups of binary noisy-OR gates. 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.

The model is based on medical expert knowledge, acquired from the medical literature and textbooks by the medical specialists in our project team. The expert knowledge is stored in a database, in such a way that extension and maintenance of the expert knowledge is easily facilitated. The database specifies the connections between variables as well as the model parameters. The parameters of a noisy-OR gate are the coupling strengths and the leak. The coupling strength to a certain disease is the probability that a finding occurs when only that disease is present and all other causes are absent. This is closely related to the sensitivity of a finding in relation to a disease when prior disease probabilities are low. The leak is the probability that the finding occurs when all the modeled diseases are absent. Prevalences range from 0.001 to 0.1. From this database, the graphical model and a user-interface for Promedas are automatically compiled (see fig. 2). This automatic procedure greatly facilitates changes in the model, such as adding or removing diseases, as required in the design phase of the model.

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.

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. Such computation can become prohibitive in particular due to the large number of parents that nodes can have. Before computation, we remove all unclamped findings from the graph, and we absorb negative findings in the prevalences [8]. Thus, only a network of positively clamped findings and their parents remain.

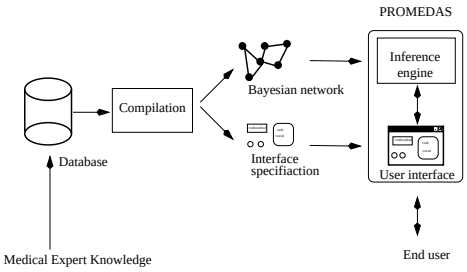

Fig. 1. Organization of PROMEDAS development.

In addition, we use efficient implementation of noisy-OR relations as proposed by [12] to reduce the size of these tables.

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.

3 Inference in the graphical model

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. The symptoms themselves, or findings $f_i \in \{0,1\}$, $i \in \{1,\ldots,N_{\mathrm{F}}\}$ constitute the bottom layer. 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.

Once a graphical model has been generated based on the above definitions, we use Bayesian inference to compute the probability of all diagnoses in the model given the patient data. Such computation can become prohibitive in particular due to the large number of parents that nodes can have. Before computation, we remove all unclamped findings from the graph, and we absorb negative findings in the prevalences [8]. Thus, only a network of positively clamped findings and their parents remain. In addition, we use efficient implementation of noisy-OR relations as proposed by [12] to reduce the size of these tables.

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.

3.1 Reduction of noisy-or clique-size

Using standard techniques for the calculation of posterior distributions directly on the factor graph in the above representation, either with a junction tree algorithm or approximation techniques, is limited to cases in which the 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].

We have illustrated the reduction of the factor potential sizes using this principle in figure 2. Note that there are many possibilities for different OR-chain constructions for a given number of dummy variables. In principle, a different choice may lead to different complexity of the resulting graph in terms of treewidth for the junction tree algorithm, in particular when diagnoses are implicated in multiple positive findings. We have distinguished between two general strategies, the first corresponding to the sequential addition of new diagnosis variables to a chain of OR-potentials (middle example in figure 2), the second corresponding to the addition of a new dummy for each pair of diagnosis parents, forming several branches of a tree (right in figure 2). Since the first construction always led to smaller treewidths in the cases under study, we chose this strategy in the graph construction.

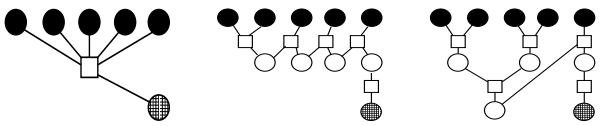

Fig. 2. Two alternatives (middle, right) for adaptation of the original graphical structure (left) by adding dummy nodes (white circles). The factor potentials (white squares) connect to at most 3 nodes in the bottom graphs. Black circles are the diagnosis nodes in the parent set of one finding node (grey circle).

3.2 Approximate inference techniques

After the reduction of the factor sizes by addition of dummy variables in one of the ways described above, we may apply a junction tree algorithm [14], or approximate inference algorithms. The complexity of the problem now depends on the set of findings that is given as input. Especially in cases where findings

share more than one common possible diagnosis, and consequently loops occur, the model can become complex. The treewidth of the junction tree, i.e. the maximal clique size, is a relevant measure for the computational complexity of the problem. When this approaches 30, the necessary exact summations over all states within the clique (in this case 2^{30} states) become problematic, and approximate inference remains the only option.

The Bethe Approximation The Bethe approximation works particularly well in cases where the graphical structure is close to treelike, since on a tree the underlying hypothesis is true. This major underlying hypothesis is factorization of the probability distribution over the Markov blanket of every node in the graph when this node itself is absent.

The marginal probabilities produced by BP, when it converges, are often a good approximation of the true marginals, and since the running time scales roughly linearly with the system size on sparse graphs, it is of particular interest for Promedas-type graphical models.

In our experiments, we added a preprocessing stage to the graph constructing routine, where we merged potential factors sharing more than one common variable node into superpotentials. Hereby a number of short loops that lead to bad performance of BP are eliminated.

Loop Corrected Belief Propagation Recently, the idea to construct schemes that improve the Bethe approximation based on expansion techniques has resulted in various new methods [15–17]. The common factor in all these approaches is that the correction terms are in some way associated to loops in the graph, that cause violations of the tree Ansatz.

Whenever a graph is not a tree, the removal of a node does not necessarily decouple the neighbours of that node. On the contrary, loops in the graph connect these neighbour nodes and cause nontrivial correlations between them even in the absence of the central node. When the interaction between these variables along the loop is small, or the loops are long, then the correlations die out, and the Bethe approximation is relatively accurate. However, the accuracy may be enhanced by taking into account these correlations, for which an estimate can be obtained with the Bethe approximation itself. The specific approach we adopt in this paper, the Loop Corrected BP (LCBP) algorithm presented in [11], is based on this idea, by Rizzo and Montanari [15]. It amounts to an easily implemented algorithm, without the need of detailed knowledge of the graph structure.

Since a detailed description of the algorithm is beyond the scope of this article, we refer to [11].

4 Simulations with virtual patient data

The Promedas model consists of many diagnosis nodes and corresponding findings. Since the relation between the two is given in terms of the probability

that a certain finding is true given that the patient has a disease, the model itself is very suitable for generating virtual patient cases. All our current results are based on such virtual patient data, which are generated as follows: first we randomly (uniformly) choose a fixed number N_d of diseases to be set to the true value. Each disease induces a set of true clamped findings according to the probability given by the model parameters. This mimics the situation in which a patient with N_d diseases reports findings that are abnormal only. With these virtual patient-cases we neglect the effect of negative findings, which affect the disease probability, but do not contribute to the computational complexity (for the computation their effect may be absorbed in the priors). The marginal disease probabilities for this "patient" calculated with our approximate inference algorithm may subsequently be compared to exact results obtained with the junction tree algorithm, as long as the latter are available. Note that the number of diagnoses that needs to be considered for each patient is typically much larger than the number of diagnoses N_d that generated the findings. The complexity of the inference task thus depends on the patient case, and in particular on the number of positive findings. Empirically we find that exact inference (the junction tree method) can be applied to patient cases with less than 30 positive findings, but this number may be significantly smaller, depending on the patient case. It is therefore interesting to investigate whether approximate methods such as BP provide a reliable alternative.

4.1 Belief Propagation results

We generated, 1000 patient cases with $N_d = 1$ and another 1000 with $N_d = 4$.The first, rather surprising, 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 error in the BP results when compared with the exact JT method are shown in fig. 3. Errors are typically small but may occasionally be rather large (fig. 3 left). In fig. 3 right 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 3 right we conclude that the quality of the BP approximation is only mildly dependent on this complexity. It is thus expected that the BP error for complex patient cases for which we cannot compute the exact result are equally reliable. We thus conclude that for patient cases where exact computation is infeasible, BP gives a reliable alternative for most cases.For those cases where the error is unacceptably large we propose to use the so-called loop corrected BP method.

4.2 Loop correction results

In the right picture of figure 4, we have plotted 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 for which the error is nonzero (i.e. loopy

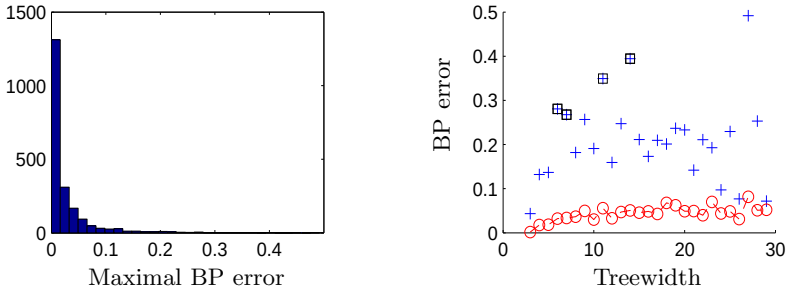

Fig. 3. *Left*: Histograms of the maximal BP single node marginal errors *Right* : BP maximal error($\circ$) averaged over instances, maximum maximal error (+) as a function of treewidth for $N_d = 4$. The squares mark instances which we have subjected to LCBP (see table below).

graphs) are plotted, 86 in total. Clearly the maximal error in the marginals produced by BP is usually about one order of magnitude larger than is the case after LCBP. The maximal maximum over all cases in this sample reduced from 0.275 to 0.023.

As a second test, we applied the method to a few cases in the right picture of figure 3, where we attempted to reduce the largest BP errors of these complex multiple disease errors. A drawback of LCBP is its rather large computation time. The computation time grows as N^2 (assuming constant maximal degree per node), and grows exponentially in the number of nodes in the largest Markov blanket in the implementation we used, since for each state in the Markov blanket of each node, one has to run BP once. The exponential scaling of the algorithm in Markov blanket size forced us to look at a few relatively easy cases only. Results for the points marked by a black square in figure 3 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	$T > T_{\max}$	$T > T_{\max}$
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. Clearly, in this form, the algorithm is not suitable for graphical models with large Markov blankets. The solution to this problem is an alternative implementation, that takes into account only nontrivial correlations between pairs of variables in the Markov blanket

(see [15]), and would consequently scale 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.

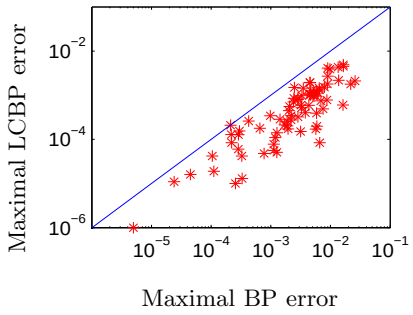

Fig. 4. maximal single node marginal error of LCBP (vertical) versus BP (horizontal). All data lie on the side of the line where the LCBP error is smaller than the BP error.

5 Conclusions and discussion

In this paper we have shown that BP is an attractive alternative 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. However, our current LCBP implementation should be improved before this method can be used in practice. One such improvement was proposed in [15].

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.

Acknowledgements The research reported here is part of the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry

of Economic Affairs, grant BSIK03024. B.W. Acknowledges the Dutch Foundation for Technology (STW) for funding.

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. S. Jaakkola and M. I. Jordan. Variational probabilistic inference and the QMR-DT network. *Journal of Artificial Intelligence Research*, 10:291, 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, T Rizzo, and H. J. Kappen. *to appear in Proceedings of AISTATS 2007*, 2007.
12. 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.
13. D. Heckerman. A tractable inference algorithm for diagnosing multiple diseases. In *Proceedings UAI*, pages 163–171. Elsevier Science, 1989.
14. 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.
15. A. Montanari and T. Rizzo. How to compute loop corrections to the Bethe approximation. *Journal of Statistical Mechanics*, page P10011, 2005.
16. J. Parisi and F. Slanina. Loop expansion around the Bethe-Peierls approximation for lattice models. *Journal of Statistical Mechanics*, page L02003, 2006.
17. M. Chertkov and V. Y. Chernyak. Loop series for discrete statistical models on graphs. *preprint cond-mat/0603189*, 2006.