THE CLUSTER VARIATION METHOD FOR APPROXIMATE REASONING IN MEDICAL DIAGNOSIS

H.J. KAPPEN
Laboratory of Biophysics, University of Nijmegen,
bert@mbfys.kun.nl

In this paper, we discuss the rule based and probabilistic approaches to computer aided medical diagnosis. We conclude that the probabilistic approach is superior to the rule based approach, but due to its intractability, it requires approximations for large scale applications. Subsequently, we review the Cluster Variation Method and derive a message passing scheme that is efficient for large directed and undirected graphical models. When the method converges, it gives close to optimal results.

1 Introduction

Medical diagnosis is the 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.

The use of computer programs to aid in the diagnostic process has been a long term goal of research in artificial intelligence. Arguably, it is the most typical application of artificial intelligence.

The different systems that have been developed so-far use a variety of modeling approaches which can be roughly divided into two categories: rule-based approaches with or without uncertainty and probabilistic methods. The rule-based systems can be viewed as computer implementations of the protocols, as described above. They consist of a large data base of rules of the form: $A \rightarrow B$, meaning that "if condition A is true, then perform action B"

or "if condition A is true, then condition B is also true". The rules may be deterministic, in which case they are always true, or 'fuzzy' in which case they are true to a (numerically specified) degree. Examples of such programs are Meditel[1], Quick Medical Reference (QMR)[2], DXplain[3], and Iliad[4].

In Berner et al.[5] a detailed study was reported that assesses the performance of these systems. A panel of medical experts collected 110 patient cases, and concensus was reached on the correct diagnosis for each of these patients. For each disease, there typically exists a highly specific test that will unambiguously identify the disease. Therefore, based on such complete data, diagnosis is easy. A more challenging task was defined by removing this defining test from each of the patient cases. The patient cases were presented to the above 4 systems. Each system generated its own ordered list of most likely diseases. In only 10-20 % of the cases, the correct diagnosis appeared on the top of these lists and in approximately 50 % of the cases the correct diagnosis appeared in the top 20 list. Many diagnoses that appeared in the top 20 list were considered irrelevant by the experts. It was concluded that these systems are not suitable for use in clinical practice.

There are two reasons for the poor performance of the rule based systems. One is that the rules that need to be implemented are very complex in the sense that the precondition A above is a conjunction of many factors. If each of these factors can be true or false, there is a combinatoric explosion of conditions that need to be described. It is difficult, if not impossible, to correctly describe all these conditions. The second reason is that evidence is often not deterministic (true or false) but rather probabilistic (likely or unlikely). The above systems provide no principled approach for the combination of such uncertain sources of information.

A very different approach is to use probability theory. In this case, one does not model the decision tree directly, but instead models the relations between diseases and symptoms in one large probability model. As a (too) simplified example, consider a medical domain with a number of diseases $d = (d_1, \ldots, d_n)$ and a number of symptoms or findings $f = (f_1, \ldots, f_m)$. One estimates the probability of each of the diseases $p(d_i)$ as well as the probability of each of the findings *given* a disease, $p(f_j|d_i)$. If diseases are independent, and if findings are conditionally independent given the disease, the joint probability model is given by:

$$p(d, f) = p(d)p(f|d) = \prod_i p(d_i) \prod_j p(f_j|d_i) \tag{1}$$

It is now possible to compute the probability of a disease d_i, given some

findings by using Bayes' rule:

$$p(d_i|f_t) = \frac{p(d_i, f_t)}{p(f_t)}, \tag{2}$$

where f_t is the list of findings that has been measured up to diagnostic iteration t. Computing this for different d_i gives the list of most probable diseases given the current findings f_t and provides the tentative diagnosis of the patient. Furthermore, one can compute which additional test is expected to be most informative about any one of the diagnoses, say d_i, by computing

$$I_{ij} = -\sum_{f_j} p(f_j|f_t) \sum_{d_i} p(d_i|f_t, f_j) \log p(d_i|f_t, f_j)$$

for each test j that has not been measured so-far. The test j that minimizes I_{ij} is the most informative test, since averaged over its possible outcomes, it gives the distribution over d_i with the lowest entropy.

Thus, one sees that whereas the rule based systems model the diagnostic process directly, the probabilistic approach models the relations between diseases and findings. The diagnostic decisions (which test to measure next) is then computed from this model. The advantage of this latter approach is that the model is much more transparent about the medical knowledge, which facilitates maintenance (changing probability tables, adding diseases or findings), as well as evaluation by external experts.

One of the main drawbacks of the probabilistic approach is that it is intractable for large systems. The computation of marginal probabilities requires summation over all other variables. For instance, in Eq. 2

$$p(f_t) = \sum_{d,f} \delta_{f,f_t} p(d, f)$$

and the sum over d, f contains exponentially many terms. Therefore, probabilistic models for medical diagnosis have been restricted to very small domains[6,7] or when covering a large domain, at the expense of the level of detail at which the disease areas are modeled[8].

In order to make the probabilistic approach feasible for large applications one therefore needs to make approximations. One can use Monte Carlo sampling but one finds that accurate results require very many iterations. An alternative is to use analytical approximations such as for instance mean field theory[9,10]. This approach works well for probability distributions that resemble spin systems (so-called Boltzmann Machines) but, as we will see, they perform poorly for directed probability distributions of the form Eq. 1.

2 The Cluster Variation Method

A very recent development is the application of the Cluster Variation method (CVM) to probabilistic inference. CVM is a method that has been developed in the physics community to approximately compute the properties of the Ising model[11]. The CVM approximates the probability distribution by a number of (overlapping) marginal distributions (clusters). The quality of the approximation is determined by the size and number of clusters. When the clusters consist of only two variables, the method is known as the Bethe approximation. Recently, the method has been introduced by Yedidia et al.[12] into the machine learning community, showing that in the Bethe approximation, the CVM solution coincides with the fixed points of the belief propagation algorithm. Belief propagation is a message passing scheme, which is known to yield exact inference in tree structured graphical models[13]. However, BP can can also give impressive results for graphs that are not trees[14].

Let $x = (x_1, \ldots, x_n)$ be a set of variables, where each x_i can take a finite number of values. Consider a probability distribution on x of the form

$$p_H(x) = \frac{1}{Z(H)} e^{-H(x)} \qquad Z = \sum_x e^{-H(x)}$$

It is well known, that p_H can be obtained as the minimum of the free energy, which is a functional over probability distributions of the following form:

$$F_H(p) = \langle H \rangle + \langle \log p \rangle \,, \tag{3}$$

where the expectation value is taken with respect to the distribution p, i.e. $\langle H \rangle = \sum_x p(x) H(x)$. When one minimizes $F_H(p)$ with respect to p under the constraint of normalization $\sum_x p(x) = 1$, one obtains p_H [a].

Computing marginals of p_H such as $p_H(x_i)$ or $p_H(x_i, x_j)$ involves sums over all states, which is intractable for large n. Therefore, one needs tractable approximations to p_H. The cluster variation method replaces the probability distribution $p_H(x)$ by a large number of (possibly overlapping) probability distributions, each describing the interaction between a small number of variables. Due to the one-to-one correspondence between a probability distribution and the minima of a free energy we can define approximate probability distributions by constructing approximate free energies and computing their minimum (or minima!). This is achieved by approximating Eq. 3 in terms of the cluster probabilities. The solution is obtained by minimizing this approximate free energy subject to normalization and consistency constraints.

[a] Minimizing the free energy can also be viewed as maximizing the entropy with an additional constraint on $\langle H \rangle$.

Define clusters as subsets of distinct variables: $x_\alpha = (x_{i_1}, \ldots, x_{i_k})$, with $1 \leq i_j \leq n$. Define a set of clusters P that contain the interactions in H and write H as a sum of these interactions:

$$H(x) = \sum_{\alpha \in P} H^\dagger_\alpha(x_\alpha)$$

For instance for Boltzmann-Gibbs distributions, $H(x) = \sum_{i>j} w_{ij} x_i x_j + \sum_i \theta_i x_i$ and P consists of all pairs and all singletons: $P = \{\alpha | \alpha = (ij), i > j \text{ or } \alpha = (i)\}$. For directed graphical models with evidence, such as Eq. 2, P is the set of clusters formed by each node i and its parent set π_i: $P = \{\alpha | \alpha = (i, \pi_i), i = 1, \ldots, n\}$. x is the set of non-evidence variables (d in this case) and $Z = p(f_t)$.

We now define a set of clusters B, that will determine our approximation in the cluster variation method. B should at least contain the interactions in $p(x)$ in the following way:

$$\forall \alpha \in P \Rightarrow \exists \alpha' \in B, \alpha \subset \alpha'.$$

In addition, we demand that no two clusters in B contain each other: $\alpha, \alpha' \in B \Rightarrow \alpha \not\subset \alpha', \alpha' \not\subset \alpha$. Clearly, the minimal choice for B is to chose clusters from P itself. The maximal choice for B is the cliques obtained when constructing the junction tree[15]. In this case, the clusters in B form a tree structure and the CVM method is exact. In general, one, can chose any set of clusters B that satisfy the above definition. Since the proposed method scales exponentially in the size of the clusters in B, the smaller the clusters in B, the faster the approximation. For a simple directed graphical model an intermediate choice of clusters is illustrated in Fig. 1.

Define a set of clusters M that consist of any intersection of a number of clusters of B: $M = \{\beta | \beta = \cap_k \alpha_k, \alpha_k \in B\}$, and define $U = B \cup M$. Once U is given, we define numbers a_β recursively by the Moebius formula

$$1 = \sum_{\alpha \in U, \alpha \supset \beta} a_\alpha, \quad \forall \beta \in U$$

In particular, this shows that $a_\alpha = 1$ for $\alpha \in B$.

The Moebius formula allows us to rewrite interactions on potentials in P in terms of interactions on clusters in U:

$$H(x) = \sum_{\beta \in P} H^\dagger_\beta(x_\beta) = \sum_{\beta \in P} \sum_{\alpha \in U, \alpha \supset \beta} a_\alpha H^\dagger_\beta(x_\beta) = \sum_{\alpha \in U} a_\alpha H_\alpha,$$

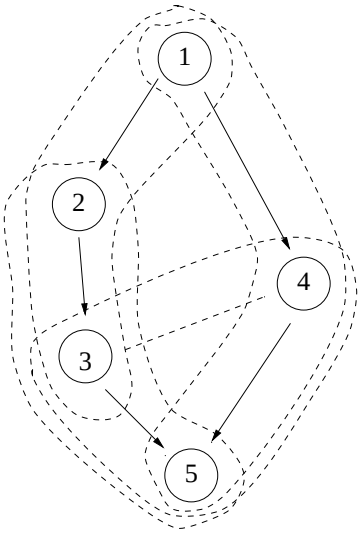


Figure 1. Directed graphical model consisting of 5 variables. Interactions are defined on clusters in $P = \{(1), (1,2), (2,3), (1,4), (3,4,5)\}$. The clusters in B are depicted by the dashed lines ($B = \{(1,2,3), (2,3,5), (1,4,5), (3,4,5)\}$. The set $M = \{(1), (2,3), (3), (5), (3,5)\}$

where we have defined H_α as the sum of all interactions in $\beta \in P$ that are contained in cluster $\alpha \in U$:

$$H_\alpha(x_\alpha) = \sum_{\beta \in P, \beta \subset \alpha} H^\dagger_\beta(x_\beta)$$

Since interactions may appear in multiple clusters, the constants a_α ensure that double counting is compensated for. [b] Thus, we can express $\langle H \rangle$ in Eq. 3 explicitly in terms of the cluster probabilities p_α as

$$\langle H \rangle = \sum_{\alpha \in U} a_\alpha \langle H_\alpha \rangle = \sum_{\alpha \in U} a_\alpha \sum_{x_\alpha} H_\alpha(x_\alpha) p_\alpha(x_\alpha) \tag{4}$$

[b] In the case of the Boltzmann distribution

$$\begin{aligned} H^\dagger_i &= H_i = \theta_i x_i \\ H^\dagger_{ij} &= w_{ij} x_i x_j \\ H_{ij} &= w_{ij} x_i x_j + \theta_i x_i + \theta_j x_j \end{aligned}$$

and $a_{(ij)} = 1$ and $a_{(i)} = 2 - n$.

Whereas $\langle H \rangle$ can be written exactly in terms of p_α, this is not the case for the entropy term in Eq. 3. The approach is to decompose the entropy of a cluster α in terms of 'connected entropies' in the following way: [c]

$$S_\alpha = -\sum_{x_\alpha} p_\alpha(x_\alpha) \log p_\alpha(x_\alpha) = \sum_{\beta \subset \alpha} S^\dagger_\beta. \tag{5}$$

Such a decomposition can be made for any cluster. In particular it can be made for the 'cluster' consisting of all variables, so that we obtain

$$S = -\sum_{x} p(x) \log p(x) = \sum_{\beta} S^\dagger_\beta \tag{6}$$

where β runs over all subsets of variables [d]. The cluster variation method approximates the total entropy by restricting this sum to only clusters in U and re-expressing $S^\dagger_\beta$ in terms of S_α, using the Moebius formula and the definition Eq. 5.

$$S \approx \sum_{\beta \in U} S^\dagger_\beta = \sum_{\beta \in U} \sum_{\alpha \supset \beta} a_\alpha S^\dagger_\beta = \sum_{\alpha \in U} a_\alpha S_\alpha \tag{7}$$

Since S_α is a function of p_α (Eq. 5) we have expressed the entropy in terms of cluster probabilities p_α.

The quality of this approximation is illustrated in Fig. 2. Note, that the both the Bethe and Kikuchi approximation strongly deteriorate around $J = 1$, which is where the spin-glass phase starts. For $J < 1$, the Kikuchi approximation is superior to the Bethe approximation. Note, however, that this figure only illustrates the quality of the truncations in Eq. 7 assuming that the exact marginals are known. It does not say anything about the accuracy of the approximate marginals using the approximate free energy.

Substituting Eqs. 4 and 7 into the free energy Eq. 3 we obtain the approximate free energy of the Cluster Variation method. This free energy must be minimized subject to normalization constraints $\sum_{x_\alpha} p_\alpha(x_\alpha) = 1$ and consistency constraints

$$p_\alpha(x_\beta) = p_\beta(x_\beta), \quad \beta \in M, \alpha \in B, \beta \subset \alpha. \tag{8}$$

Note, that we have excluded constraints between clusters in M. This is sufficient because when $\beta, \beta' \in M$, $\beta \subset \beta'$ and $\beta' \subset \alpha \in B$: $p_\alpha(x_{\beta'}) = p_{\beta'}(x_{\beta'})$

[c] This decomposition is similar to writing a correlation in terms of means and covariance. For instance when $\alpha = (i)$, $S_{(i)} = S^\dagger_{(i)}$ is the usual mean field entropy and $S_{(ij)} = S^\dagger_{(i)} + S^\dagger_{(j)} + S^\dagger_{(ij)}$ defines two node correction.

[d] On n variables this sum contains 2^n terms.

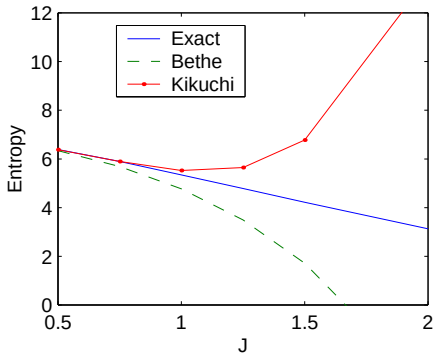


Figure 2. Exact and approximate entropies for the fully connected Boltzmann-Gibbs distribution on $n = 10$ variables with random couplings (SK model) as a function of mean coupling strength. Couplings w_{ij} are chosen from a normal Gaussian distribution with mean zero and standard deviation $J/\sqrt{n}$. External fields θ_i are chosen from a normal Gaussian distribution with mean zero and standard deviation 0.1. The exact entropy is computed from Eq. 6. The Bethe and Kikuchi approximations are computed using the approximate entropy expression Eq. 7 with exact marginals and by choosing B as the set of all pairs and all triplets, respectively.

and $p_\alpha(x_\beta) = p_\beta(x_\beta)$ implies $p_{\beta'}(x_\beta) = p_\beta(x_\beta)$. In the following, α and β will be from B and M respectively, unless otherwise stated [e].

Adding Lagrange multipliers for the constraints we obtain the Cluster Variation free energy:

$$F_{\rm cvm}(\{p_\alpha(x_\alpha)\},\{\lambda_\alpha\},\{\lambda_{\alpha\beta}(x_\beta)\}) = \sum_{\alpha\in U} a_\alpha \sum_{x_\alpha} p_\alpha(x_\alpha)\left(H_\alpha(x_\alpha) + \log p_\alpha(x_\alpha)\right)$$
$$- \sum_{\alpha\in U}\lambda_\alpha\left(\sum_{x_\alpha} p_\alpha(x_\alpha) - 1\right) - \sum_{\alpha\in U}\sum_{\beta\subset\alpha}\sum_{x_\beta}\lambda_{\alpha\beta}(x_\beta)\left(p_\alpha(x_\beta) - p_\beta(x_\beta)\right) \tag{9}$$

3 Iterating Lagrange multipliers

Since the Moebius numbers can have arbitrary sign, Eq. 9 consists of a sum of convex and concave terms, and therefore is a non-convex optimization problem. One can separate $F_{\rm cvm}$ in a convex and concave term and derive an

[e] In fact, additional constraints can be removed, when clusters in M contain subclusters in M. See Kappen and Wiegerinck[16].

iteration procedure in p_α and the Lagrange multipliers that is guaranteed to converge[17]. The resulting algorithm is a 'double loop' iteration procedure.

Alternatively, by setting $\frac{\partial F_{\rm cvm}}{\partial p_\alpha(x_\alpha)}, \alpha \in U$ equal to zero, one can express the cluster probabilities in terms of the Lagrange multipliers:

$$p_\alpha(x_\alpha) = \frac{1}{Z_\alpha} \exp\left(-H_\alpha(x_\alpha) + \sum_{\beta \subset \alpha} \lambda_{\alpha\beta}(x_\beta)\right) \tag{10}$$

$$p_\beta(x_\beta) = \frac{1}{Z_\beta} \exp\left(-H_\beta(x_\beta) - \frac{1}{a_\beta} \sum_{\alpha \supset \beta} \lambda_{\alpha\beta}(x_\beta)\right) \tag{11}$$

The remaining task is to solve for the Lagrange multipliers such that all constraints (Eq. 8) are satisfied. There are two ways to do this. One is to define an auxiliary cost function that is zero when all constraints are satisfied and positive otherwise and minimize this cost function with respect to the Lagrange multipliers. This method is discussed in Kappen and Wiegerinck[16].

Alternatively, one can substitute Eqs. 10-11 into the constraint Eqs. 8 and obtain a system of coupled non-linear equations. In Yedidia et al.[12] a message passing algorithm was proposed to find a solution to this problem. Here, we will present an alternative method, that solves directly in terms of the Lagrange multipliers.

Consider the constraints Eq. 8 for some fixed cluster β and all clusters $\alpha \supset \beta$ and define $B_\beta = \{\alpha \in B | \alpha \supset \beta\}$. We wish to solve for all constraints $\alpha \supset \beta$, with $\alpha \in B_\beta$ by adjusting $\lambda_{\alpha\beta}, \alpha \in B_\beta$. This is a sub-problem with $|B_\beta||x_\beta|$ equations and an equal number of unknowns, where $|B_\beta|$ is the number of elements of B_β and $|x_\beta|$ is the number of values that x_β can take. The probability distribution p_β (Eq. 11) depends only on these Lagrange multipliers, up to normalization. p_α (Eq. 10) depends also on other Lagrange multipliers. However, we consider only its dependence on $\lambda_{\alpha\beta}, \alpha \in B_\beta$, and consider all other Lagrange multipliers as fixed. Thus,

$$p_\alpha(x_\alpha) = \exp(\lambda_{\alpha\beta}(x_\beta))\tilde{p}_\alpha(x_\alpha), \alpha \in B_\beta \tag{12}$$

with $\tilde{p}_\alpha$ independent of $\lambda_{\alpha\beta}, \alpha \in B_\beta$.

Substituting, Eqs. 11 and 12 into Eq. 8, we obtain a set of linear equations for $\lambda_{\alpha\beta}(x_\beta)$ which we can solve in closed form:

$$\lambda_{\alpha\beta}(x_\beta) = -\frac{a_\beta}{a_\beta + |B_\beta|} H_\beta(x_\beta) - \sum_{\alpha'} A_{\alpha\alpha'} \log \tilde{p}_{\alpha'}(x_\beta)$$

with

$$A_{\alpha\alpha'} = \delta_{\alpha\alpha'} - \frac{1}{a_\beta + |B_\beta|}$$

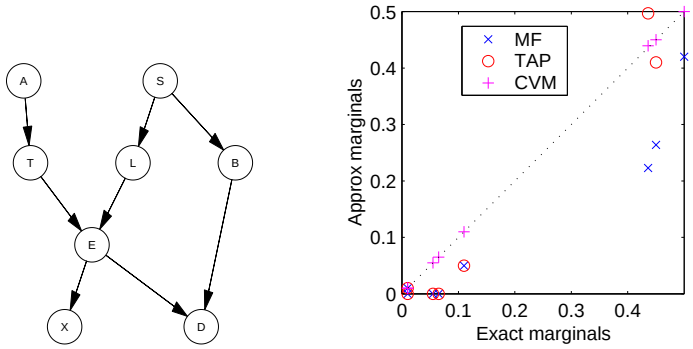
We update the probabilities with the new values of the Lagrange multipliers using Eqs. 11 and 12. We repeat the above procedure for all $\beta \in M$ until convergence.

4 Numerical results

We show the performance of the Lagrange multiplier iteration method (LMI) on several 'real world' directed graphical models. For undirected models, see Kappen and Wiegerinck[16]. First, we consider the well-known chest clinic problem, introduced by Lauritzen and Spiegelhalter[15]. The graphical model is given in figure 3a. The model describes the relations between three diagnoses (Tuberculosis(T), Lung Cancer(L) and Bronchitis(B), middle layer), clinical observations and symptoms (Positive X-ray(X) and Dyspnoea(D)(=shortness of breath), lower layer) and prior conditions (recent visit to Asia(A) and whether the patient smokes(S)). In figure 3b, we plot the exact single node marginals against the approximate marginals for this problem. For LMI, the clusters in B are defined according to the conditional probability tables, i.e. when a node has k parents, a cluster of size $k+1$ on this node and its parents is included in the set B. Convergence was reached in 6 iterations. Maximal error on the marginals is 0.0033. For comparison, we computed the mean field and TAP approximations, as previously introduced by Kappen and Wiegerinck[10]. Although TAP is significantly better than MF, it is far worse than the CVM method. This is not surprising, since both the MF and TAP approximation are based on single node approximation, whereas the CVM method uses potentials up to size 3.

Secondly, we consider a graphical model that was developed in a project together with the department of internal medicine of the Utrecht Academic hospital. In this project, called Promedas, we aim to model a large part of internal medicine[18]. The network that we consider was one of the first modules that we built and models in detail some specific anemias and consists of 91 variables. The network was developed using our graphical tool BayesBuilder[19] which is shown with part of the network in figure 4. The clusters in B are defined according to the conditional probability tables. Convergence was reached in 5 iterations. Maximal absolute error on the marginals is 0.0008. The mean field and TAP methods perform very poorly on this problem.

Finally, we tested the cluster variation method on randomly generated



(a) Chest clinic graphical model

(b) Approximate inference

Figure 3. a) The Chest Clinic model describes the relations between diagnoses, findings and prior conditions for a small medical domain. An arrow $a \rightarrow b$ indicates that the probability of b depends on the values of a. b) Inference of single node marginals using MF, TAP and LMI method, comparing the results with exact.

directed graphical models. Each node is randomly connected to k parents. The entries of the probability tables are randomly generated between zero and one. Due to the large number of loops in the graph, the exact method requires exponential time in the so-called tree width, which can be seen from Table 1 to scale approximately linear with the network size. Therefore exact computation is only feasible for small graphs (up to size $n = 40$ in this case).

For the CVM, clusters in B are defined according to the conditional probability tables. Therefore, maximal cluster size is $k + 1$. On these more challenging cases, LMI does not converge. The results shown are obtained with the auxiliary cost function as that was briefly mentioned in section 3 and fully described in Kappen and Wiegerinck[16]. Minimization was done using conjugate gradient descent. The results are shown in Table 1.

5 Conclusion

In this paper, we have described two approaches to computer aided medical diagnosis. The rule based approach directly models the diagnostic decision tree. We have shown that this approach fails to pass the test of clinical

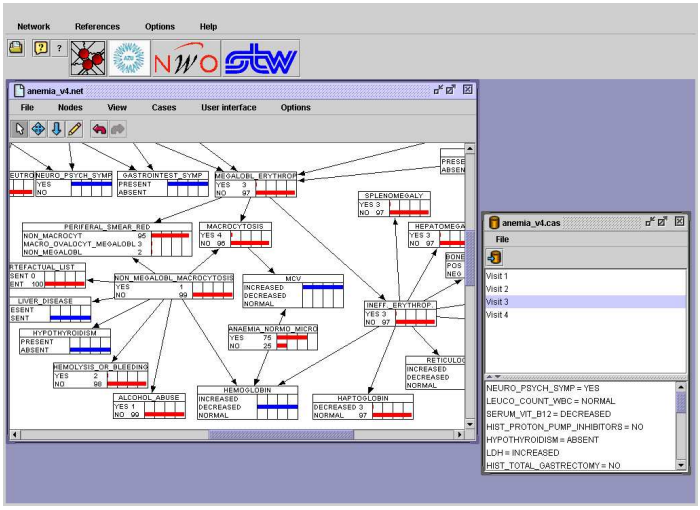


Figure 4. BayesBuilder graphical software environment, showing part of the Anemia network. The network consists of 91 variables and models some specific Anemias.

n	Iter	$\|C\|$	Potential error	Margin error	Constraint error
10	16	8	0.068	0.068	5.8e-3
20	30	12	0.068	0.216	6.2e-3
30	44	16	0.079	0.222	4.5e-3
40	48	21	0.073	0.218	4.2e-3
50	51	26	–	–	3.2e-3

Table 1. Comparison of CVM method for large directed graphical models. Each node is connected to $k = 5$ parents. $|C|$ is the tree width of the triangulated graph required for the exact computation. Iter is the number of conjugate gradient descent iterations of the CVM method. Potential error and margin error are the maximum absolute error in any of the cluster probabilities and single variable marginals computed with CVM, respectively. Constraint error is the maximum absolute error in any of the constraints Eq. 8 after termination of CVM.

relevance and we have given several reasons that could account for this failure.

The alternative approach uses a probabilistic model to describe the relations between diagnoses and findings. This approach has the great advantage that it provides a principled approach for the combination of different sources

of uncertainty. The price that we have to pay for this luxury is that probabilistic inference is intractable for large systems.

As a generic approximation method, we have introduced the Cluster Variation method and presented a novel iteration scheme, called Lagrange Multiplier Iteration. When it converges, it provides very good results and is very fast. However, it is not guaranteed to converge in general. In those more complex cases one must resort to more expensive methods, such as CCCP [17] or using an auxiliary cost function[16].

Acknowledgments

This research was supported in part by the Dutch Technology Foundation (STW). I would like to thank Taylan Cemgil for providing his Matlab graphical models toolkit, and Wim Wiegerinck and Sebino Stramaglia (Bari, Italy) for useful discussions.

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. E.S. Berner, G.D. Webster, A.A. Shugerman, J.R. Jackson, J. Algina, A.L. Baker, E.V. Ball, C.G. Cobbs, V.W. Dennis, E.P. Frenkel, L.D. Hudson, E.L. Mancall, C.E. Racley, and O.D. Taunton. Performance of four computer-based diagnostic systems. *N-Engl-J-Med.*, 330(25):1792–6, 1994.
6. 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.
7. 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.
8. 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.
9. H.J. Kappen and F.B. Rodríguez. Efficient learning in Boltzmann Machines using linear response theory. *Neural Computation*, 10:1137–1156, 1998.
10. H.J. Kappen and W.A.J.J. Wiegerinck. Second order approximations for probability models. In Todd Leen, Tom Dietterich, Rich Caruana, and Virginia de Sa, editors, *Advances in Neural Information Processing Systems 13*, pages 238–244. MIT Press, 2001.

11. R. Kikuchi. *Physical Review*, 81:988, 1951.
12. J.S. Yedidia, W.T. Freeman, and Y. Weiss. Generalized belief propagation. In T.K. Leen, T.G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13 (Proceedings of the 2000 Conference)*, 2001. In press.
13. J. Pearl. *Probabilistic reasoning in intelligent systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, California, 1988.
14. 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.
15. 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.
16. H.J. Kappen and W. Wiegerinck. A novel iteration scheme for the cluster variation method. In T.G. Dieterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, 2002. In press.
17. A.L. Yuille and A. Rangarajan. The convex-concave principle. In T.G. Dieterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, 2002. In press.
18. W. Wiegerinck, H.J. Kappen, E.W.M.T ter Braak, W.J.P.P ter Burg, M.J. Nijman, Y.L. O, and J.P. Neijt. Approximate inference for medical diagnosis. *Pattern Recognition Letters*, 20:1231–1239, 1999.
19. B. Kappen, W. Wiegerinck, and M. Nijman. Bayesbuilder. In W. Buntine, B. Fischer, and J. Schumann, editors, *Software Support for Bayesian Analysis*. RIACS, NASA Ames Research Center, 2000.