MFDTs: Mean Field Dynamic Trees

Nicholas J Adams
Division of Informatics
University of Edinburgh, UK
nicka@dai.ed.ac.uk

Amos J Storkey
Division of Informatics
University of Edinburgh,UK
a.storkey@ed.ac.uk

Zoubin Ghahramani
Gatsby Computational Neuroscience Unit
University College London, UK
zoubin@gatsby.ucl.ac.uk

Christopher K I Williams
Division of Informatics
University of Edinburgh, UK
c.k.i.williams@ed.ac.uk

Abstract

Tree structured belief networks are attractive for image segmentation tasks. However, networks with fixed architectures are not very suitable as they lead to blocky artefacts, and led to the introduction of Dynamic Trees (DTs) in [6]. The Dynamic Trees architecture provides a prior distribution over tree structures, and in [6] simulated annealing (SA) was used to search for structures with high posterior probability. In this paper we introduce a mean field approach to inference in DTs. We find that the mean field method captures the posterior better than just using the maximum a posteriori solution found by SA.

1. Introduction

Tree structured belief nets are useful for image segmentation [1, 7]. They provide a hierarchically structured model for the different picture elements. The a priori understanding behind this choice of model stems from the fact that we want the image to be segmented into a number of different regions. We would expect those regions to correspond in some way to the objects that make up the picture. The hierarchical model seems a natural one for object representation, where higher level nodes control the distribution of a large number of leaf nodes (pixels).

Quadtree-structured belief networks provide a model of this form [1, 7]. They allow exact inference through belief propagation [3]. However quadtrees produce blocky artefacts due to the fact that two (spatially) adjacent leaf nodes might only be path-connected through a vertex far up the tree hierarchy. One way around this type of problem is to use dynamic trees [6]. This is a mixture of tree structures formed by allowing each vertex to 'choose' its parent. This reduces the blockiness problem, because any two leaf nodes can be connected at any level of hierarchy, but as the number of trees in the mixture grows exponentially with network size, exact belief propagation becomes intractable.

One approach for approximating the posterior distribution of the dynamic tree involves using the maximum a posteriori choice of tree, obtained through simulated annealing [6]. Experience has shown that the annealing process tends to be slow to converge. Here a different approach is taken. Variational methods are used to fit an approximating distribution to the true posterior. A standard technique involves the use of a factorised distribution (the mean field approach) [4, 2]. It is shown here that such an approximation is useful for dynamic trees.

Section 2 of the paper describes the theory behind the mean field approach to DTs, and experiments comparing it with other methods are described in section 3.

2. Inference in Dynamic Trees

A dynamic tree belief network is a mixture of tree structured belief networks. The model consists of two components: a prior distribution of possible tree architectures, and the conditional probabilities of each node given its parents and the tree architecture. There any many different possibilities for such components. In [6], the authors used a structure based on a modified quadtree. The nodes are arranged in a layered structure, and each node 'chooses' its parent independently from those in the layer above. The *natural* parent (that which would be chosen in a quadtree arrangement) has a higher probability of being chosen than the other possible parents. Nodes which lie adjacent to the *natural* parent on the same layer are termed the *nearest neighbours*, and similarly nodes a distance N away from the natural parent are the *N'th nearest neighbours*. Also the possibility of a node choosing to be a new root node is allowed with some small probability.

The model is used by instantiating the evidential nodes (in our case the leaf nodes) of the network. We wish to infer values for the non-evidential nodes. We also want information about the posterior distribution of the tree structures of the network. As dynamic trees no longer have the simple tree structure that the quadtree networks had, tractable inference using belief propagation is no longer feasible.

Two possible approaches to this problem are considered. The first of these involves using annealing. The second of these involves using a mean field variational approach. The annealing case has been considered in an earlier paper [6]. Here the mean field approach is introduced.

2.1. Mean field for dynamic trees

Consider an ordered set V of nodes $i = 1, 2, \ldots, n$. Consider also a set S of possible states $1, 2 \ldots, m$ of each node. Let $Z = \{z_{ij}\}$ denote the set of possible directed tree structures over these nodes, where z_{ij} is an indicator. $z_{ij} = 1$ denotes the fact that node j is the parent of node i. The ordering of the nodes means that $z_{ij} \equiv 0$ for $j \geq i$. Finally let $X = \{x_i^k\}$ represent the state of the nodes: $x_i^k = 1$ if node i is in state k, and is zero otherwise.

Given the above notation, a dynamic tree can be represented by a prior over the possible trees $P(Z)$, and a prior over the network states given a particular tree structure $P(X|Z)$. This prior over the network states is given by the conditional probability tables of the network. We assume that the prior over Z factorises: in other words each node 'chooses' a parent from a set of possible parents independently of other nodes (termed the *full-time-node-employment* prior in [6]). Hence $P(Z) = \prod_{ij} \pi_{ij}^{z_{ij}}$, where π_{ij} is the probability that node i chooses parent j. The conditional probability tables define the state transition probabilities when traversing a link between a node j and its child i, where P_{ij}^{kl} is the probability of moving from state l to state k during such a transition.

With these prior forms, the joint prior distribution can be written as

$$P(Z, X) = \prod_{i=1}^{n} \prod_{j=1}^{n} \pi_{ij}^{z_{ij}} \prod_{kl} [P_{ij}^{kl}]^{x_i^k x_j^l z_{ij}} \tag{1}$$

where the indicator variables are simply used to pick out the correct probabilities.

The nodes (vertices) are split into a set V^E and a set V^H of evidential and non-evidential (hidden) nodes respectively. Likewise the corresponding node state indicator variables are denoted by X^E and X^H respectively. The posterior distribution of the dynamic tree can then be written as $P(Z, X^H|X^E) = P(Z, X)/P(X^E)$

The mean field variational approach involves approximating this posterior distribution with a factorising distribution of the form $Q(Z)Q(X^H)$, where $Q(Z)$ is the approximating distribution over the Z variables, and $Q(X^H)$ is the approximating distribution over the non-evidential X^H. To choose good forms for the Q's the Kullback-Liebler divergence between the $Q(Z)Q(X^H)$ distribution and the true posterior should be minimised. The KL divergence is of the form

$$\begin{aligned} KL(Q||P) &= \sum_{Z, X^H} Q(Z, X^H) \log\left(\frac{Q(Z)Q(X^H)}{P(Z, X^H|X^E)}\right) \\ &= \log P(X^E) - \sum_{Z, X^H} Q(Z)Q(X^H)[\log P(Z, X) \\ &\qquad - \log Q(Z) - \log Q(X^H)] \end{aligned} \tag{2}$$

Calculating Q(Z) The procedure for optimising this KL divergence is now outlined. If $Q(X^H)$ is fixed, then $Q(Z)$ can be chosen to minimise (2). Performing such a minimisation gives $\log Q(Z) = \sum_{X^H} Q(X^H) \log P(Z, X) + const$. Substituting for P from (1) and normalising, we get

$$Q(Z) = \prod_{ij} \frac{\exp(z_{ij}\lambda_{ij})}{\sum_s \exp(\lambda_{is})} \tag{3}$$

where $\lambda_{ij} = \log \pi_{ij} + \sum_{kl} \langle x_i^k x_j^l \rangle_{Q(X^H)} \log P_{ij}^{kl}$. Hence we can explicitly calculate the optimal $Q(Z)$ for fixed $Q(X^H)$. Note that $Q(Z)$ turns out to be a factorised distribution.

Calculating Q(X) The next stage involves the minimisation of the KL divergence for $Q(X^H)$ keeping $Q(Z)$ fixed. Again substituting in for $P(Z, X)$ we get

$$\begin{aligned} KL(Q||P) &= \sum_{X^H} Q(X^H) \log Q(X^H) - \sum_{ij} \mu_{ij} [\log \pi_{ij} \\ &+ \sum_{kl} \langle x_i^k x_j^l \rangle_{Q(X^H)} \log P_{ij}^{kl}] + rest \end{aligned} \tag{4}$$

where $\mu_{ij} = \langle z_{ij} \rangle_{Q(Z)}$ and where $rest$ depends only on the form of $Q(Z)$ and is hence held constant.

To do this minimisation, further assumptions need to be made about the form of $Q(X^H)$. Here we require $Q(X^H)$ to factorise further into the mean field form: $Q(X^H) = \prod_{ik} (m_i^k)^{x_i^k}$. The m_i^k denotes the mean field probability that variable i is in state k. With this assumption, (4) can be optimised with respect to the m's using a Lagrange multiplier term $\sum_\alpha \rho_\alpha (\sum_\beta m_\alpha^\beta - 1)$.

A straightforward application of calculus gives the following iterative update for the means

$$m_s^r = \frac{\exp(\gamma_s^r)}{\sum_r \exp(\gamma_s^r)} \tag{5}$$

where

$$\gamma_s^r = \sum_j \sum_l \mu_{sj} m_j^l \log P_{sj}^{rl} + \mu_{js} m_j^l \log P_{js}^{rl} \qquad (6)$$

The whole procedure These two solutions give us all the information needed to perform an optimisation of the KL divergence. Firstly the m_{ij} are initialised to $0.5 \pm \epsilon$ (where ϵ is Gaussian noise of zero mean and variance 0.01), and the μ_{ij} calculated. A local maximum of $Q(X)$ can then be found by iteratively updating (5) asynchronously across the different nodes.

Once this has suitably converged, equation (3) can be used with $\langle x_i^k x_j^l \rangle_{Q(X)} = m_i^k m_j^l$ to calculate the new conditionally optimal $Q(Z)$ distribution. (Note that $\langle x_i^k x_j^l \rangle_{Q(X)}$ needs only to be computed for $j < i$ as $z_{ij} \equiv 0$ for $j \geq i$).

This whole process is repeated until convergence. Each step of the process reduces the KL divergence (2), and so convergence is guaranteed at a local minimum.

3. Experiments

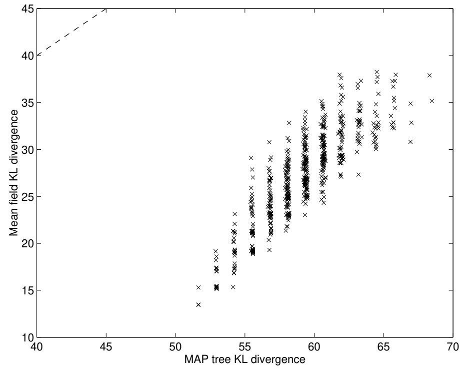
We explore and contrast the performance of the mean field approach with that of simulated annealing [6] using a 6 layer binary tree. With this architecture we have 1-d images with 32 pixels. Initially we shall consider the case where the node states are binary variables and the images are black and white.

A standard DT model of the above architecture was used. The prior over node states was set to be uniform, with conditional probabilities of 0.99 down the diagonal and 0.01 off-diagonal. The probability of nodes choosing to become a root (disconnecting) were set to be more favourable than connecting to the *nearest neighbour*, but less favourable than connecting to the *natural parent*. This was achieved in the same way as described in [6], using the prior $\pi_{ij} = \mathrm{e}^{\beta a_{ij}} / \sum_k \mathrm{e}^{\beta a_{ik}}$. The affinities, a_{ij}, were set as 1 for the *natural* parent, and $1 - N$ for the *N'th nearest neighbours* of the *natural* parent, with $\beta = 1.25$. The affinity for becoming a root, a_{null}, was 0.5. The model was sampled to generate a suite of training data of some 1000 images from which 600 were selected for our experiments.

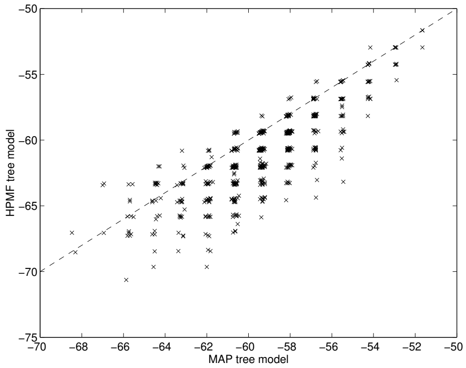
In the experiments we use simulated annealing in the same way as in [6] to find the maximum a posteriori (MAP) configuration of the DT for each of the images.

For the mean field approach we order the nodes from the bottom nodes to those on the higher levels, and sweep through them updating the ms asynchronously a total of 20 times each. This was found to be sufficient to allow these simultaneous equations to reach their equilibrium state. The $Q(Z)$s can then be recalculated. Typically the algorithm converged[1] after 4 or 5 iterations. Mean field was found to be of the order 100 times faster than simulated annealing.

To compare mean field and simulated annealing we plot the KL divergences[2] $KL(Q||P)$ against $KL(R||P)$, where R is the MAP tree configuration (see Figure 1(a)).



(a)



(b)

Figure 1. Comparison of (a) KL divergence, and (b) the unnormalised log posterior of the MAP tree against the corresponding mean field DTs.

From the comparative plot of Figure 1(a) it is clear that the KL divergence of the mean field solutions is significantly lower than that of the MAP dynamic tree in all instances. Typically we see from Figure 1(a), a difference in KL divergence of about 30 between the mean field example, and the corresponding MAP tree. These results can be understood when we realise that although the mean field approximation requires the assumption that $P(X)$ can be factorised, ie. $P(X) = \prod_i P(X_i)$, it maintains a distribution over $P(Z)$. For the MAP case we usually choose a tree with greater posterior probability, but we are only basing our estimate of the KL divergence on a single structure, which is unlikely to account for a high proportion of the probability mass of the posterior distribution. It can be seen that the distribution of points is grouped into a series of energy bands

[1] A threshold change of less than 0.1 in the KL divergence between successive steps was found to be sufficient to allow the $Q(Z)$ to stabilise on a particular configuration.

[2] The KL divergence can be computed up to the addition of a constant dependent solely on the probability of the image data, $P(X^E)$.

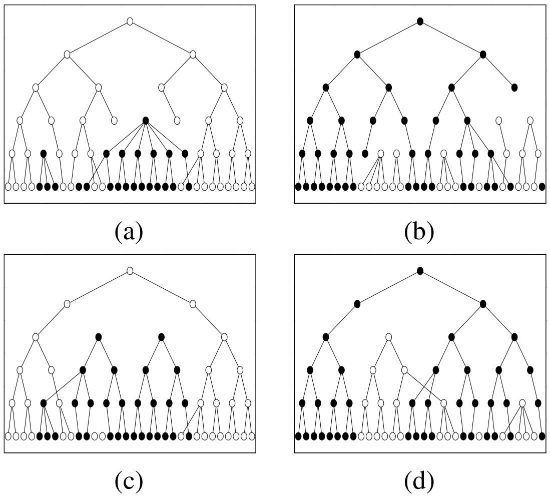


Figure 2. (a)-(b) The HPMF structures for 2 different images, and (c)-(d) the corresponding MAP trees found by annealing.

for the MAP model, whereas for the mean field method they are more evenly spread. This is probably due to the discrete nature of choices over tree structure and node state in the true posterior distribution. We plan to investigate it further.

We can also compare the posterior[3] probability of the MAP tree found by annealing and posterior of the highest probability tree structure found by mean field (HPMF tree), where the connected links in the HPMF tree are the z_{ij}s of highest probability from the $Q(Z)$ distribution. Such a comparison is made in Figure 1(b), where the line in the figure denotes the boundary of equal log posterior. We notice that in most cases (81.8%) the annealed tree has a higher posterior, but in 11.5% of cases it is the same, and for 6.7% of examples the mean field approach actually found a higher posterior tree. The latter is probably an indication that though the annealer generally finds very good optima, it cannot guarantee finding the global solution. Mean field by attempting to fit a distribution better explores the landscape and is able find some of the harder solutions. However as we cannot exactly fit the posterior we should usually expect the sampling approach to give better results. This is encouraging in that it demonstrates that the mean field is able to find interpretations of the data which are comparable in performance to the MAP structures found by sampling, at only a fraction of the computational cost.

A qualitative examination of the types of structures the mean field technique finds is quite instructive. Two examples of HPMF trees found for different images are shown in Figures 2(a) and 2(b), and the MAP trees found by simulated annealing with the same data are shown below them in Figures 2(c) and 2(d). It can be seen that there is a high degree of similarity in their structure, with both methods picking out objects in the image as separate trees.

[3] We define the posterior as $P(Z|X^E) \propto P(Z)P(X^E|Z)$ and ignore the normalising term $P(X^E)$ which is constant across the two approaches.

4. Discussion

We can conclude that the mean field approach provides significant advantages over structure searching for the MAP solution in that it produces an approximating distribution to the posterior, which is more informative than simply choosing a single example. Mean field was also able to find good HPMF solutions that rivalled the MAP structures found by simulated annealing. This was achieved with a considerable saving in computational effort and comes close to making real time inference in DTs viable.

We note, however, that the assumption in mean field of a factorised distribution over $P(X)$ is not necessarily a good one, and we hope to focus further on distributions giving a closer approximation to the true posterior. One possibility of using a tree structure to reflect their hierarchical dependence upon each other, has already been considered in [5].

Given the success of the mean field approach at finding good trees, we are now investigating learning of the DT model parameters using a mean field based procedure with a view to segmentation of real-world images.

Acknowledgements

NJA is supported by an EPSRC research studentship. The work of AS and CW is supported through EPSRC grant GR/L78161 *Probabilistic Models for Sequences*.

References

[1] C. A. Bouman and M. Shapiro. A Multliscale Random Field Model for Bayesian Image Segmentation. *IEEE Transactions on Image Processing*, 3(2):162–177, Mar. 1994.

[2] Z. Ghahramani. On Structured Variational Approximations. Technical Report CRG-TR-97-1, Department of Computer Science, University of Toronto, Canada, M5S 1A4, 1997.

[3] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufman Publishers Inc., San Francisco, USA, 1988.

[4] L. K. Saul and M. I. Jordan. Exploiting Tractable Substructures in Intractable Networks. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*. MIT Press, 1996.

[5] A. J. Storkey. Dynamic Trees: A structured Variational Approach Giving Efficient Propagation Rules. In *Uncertainty in Artificial Intelligence (UAI2000)*. 2000. To appear.

[6] C. K. I. Williams and N. J. Adams. DTs: Dynamic Trees. In M. J. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11*. MIT Press, 1999.

[7] C. K. I. Williams and X. Feng. Combining Neural Networks and Belief Networks for Image Segmentation. In *Proceedings of IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing*, 1998.