Sequential Clustering by Loopy Belief Propagation

Thomas Ott [*] Justin Dauwels [†] Ruedi Stoop [*]

Abstract — **In this contribution we exploit the potential of belief propagation in connection with sequential superparamagnetic clustering. We first give a short overview of the methods and concepts used and show how to combine them. We finally discuss some implementation issues, problems and possible advantages in comparison with Markov chain Monte Carlo methods by means of a toy system.**

1 Introduction

Clustering methods are valuable tools for the data analysis in many different scientific fields. Data structuring by clustering, i.e., the identification of 'natural' groups of similar data items, leads to a compact data representation in the form of macrostates. In many situations, no a priori information, e.g., about the number of these macrostate-clusters, is available. In these cases, we rely on non-parametric methods that are able to determine characteristics, such as the number and shape of the most natural clusters, independently. In this respect, superparamagnetic clustering (SC) [1] is a proven approach that provides 'natural' clustering solutions on different resolution levels. The levels are controlled by a 'temperature' parameter T which indirectly determines the number of clusters. The computational key task within SC is to calculate pair marginals. In the standard approach, this calculation was based on computationally expensive Markov Chain Monte Carlo methods (MCMC). However, Shental et al. [2] paved the way towards an implementation with a possibly smaller computational cost - an important issue for large data sets. They reformulated SC as a graph partitioning problem (*typical cut*) and showed that calculating the typical cut is equivalent to performing inference in graphical models. In the context of graphical models, loopy belief propagation (BP) is a well-known algorithm for an efficient calculation of marginals (see, e.g., [3]). In [2], loopy (and generalised) BP was successfully applied for obtaining an approximate solution of the typical cut problem of larger data sets.

A question of central character, however, remains: How to select the most natural resolution level and thus the number of classes among all levels provided by SC? In [4], it was demonstrated that, based on a sequential extraction of clusters, a unique (natural or stable) clustering can be found. Furthermore , it was outlined that the most natural resolution level is a local property that cannot be characterised by a global temperature T anymore. In this contribution, we exploit the potential of BP in connection with sequential clustering. The question of interest is whether BP, used for sequential clustering, outperforms the traditional MCMC approach in terms of performance and complexity.

The paper is organised as follows: For the readers new to the field , we give a brief introduction to graphical inference problems by means of pairwise Markov random fields (PMRF) and the connection to BP in Sec. 2[1]. In Sec. 3, the concept of sequential SC is reviewed and is connected to PMRFs and BP. In Sec. 4, we address some issues related to the practical implementation and explore the computational performance on the basis of a simple toy system example. Finally, Sec. 5 offers some concluding remarks.

2 Pairwise Markov Random Fields and Belief Propagation

Pairwise Markov random fields

PMRFs have become attractive models for inference problems in computer vision. In such problems, we usually want to infer some information about the visual scene from an array of pixel intensities $y_i, i = 1, ..., N$. Consider as an example the task of a figure-ground segmentation. x_i denotes to which segment pixel i belongs (e.g., background: $x_i = 0$; figure: $x_i = 1$). Normally, some *evidence* for x_i is available which can be expressed as a function $\phi_i(x_i, y_i)$. Furthermore, there usually exists a statistical dependency between the x_i-values of neighbouring (i.e. 'connected') pixels: If two neighbouring pixels show similar characteristics then they are likely to belong to the same segment. This dependency is represented by pairwise functions $\psi_{ij}(x_i, x_j)$.

For fixed y_i, the overall joint probability of a scene x_i is taken to be

$$p(\{x\},\{y\}) = \frac{1}{Z}\prod_{(i,j)}\psi_{ij}(x_i,x_j)\prod_i \phi_i(x_i,y_i). \quad (1)$$

[*] Institute of Neuroinformatics, ETH/UNIZH Zurich, Switzerland, e-mail: [tott, ruedi]@ini.phys.ethz.ch, tel.: +41 44 635 30 63

[†] Dept. of Information Technology and Electrical Engineering, ETH, CH-8092 Zürich, Switzerland, dauwels@isi.ee.ethz.ch.

[1] In Sec. 2, we essentially follow [3].

Z is a normalisation constant. Typically, we are interested in the marginals $p_i(x_i)$. In connection with SC, we are also interested particularly in the pair-marginals

$$p_{ij}(x_i, x_j) = \sum_{x_k, k \neq i,j} p(x_1, ..., x_N). \qquad (2)$$

Belief propagation

For large systems it is hopeless to carry out all the sums in (2) as the number of calculations increases exponentially with N. BP is an algorithm for an efficient calculation of these marginals. In the case of loop-free PMRFs the algorithm was shown to provide the exact marginals (in a time proportional to the number of links in the graph [3]). However, in the case of loopy fields there is no guarantee that BP does converge, and if so, that it converges to useful values. Despite this uncertainty, BP was proven to deliver useful approximations for marginals of many loopy problems. Superparamagnetic clustering poses a problem on a highly loopy graph. We will thus concentrate on a 'loopy' BP implementation.

In the language of BP, messages between connected nodes of the PMRF are interchanged. The message $m_{ij}(x_j)$ sent from node i to node j contains a recommendation about what state node j should be in (e.g., if $m_{ij}(1) < m_{ij}(0)$ then $x_j = 0$ should be preferred). Given the set of messages at time t, $\{m_{ij}^t(x_j)\}$, the messages at time $t+1$ are determined by

$$m_{ij}^{t+1}(x_j) = \sum_{x_i} \phi_i(x_i) \psi_{ij}(x_i, x_j) \prod_{k \in N(i) \setminus j} m_{ki}^t(x_i). \qquad (3)$$

$N(i) \setminus j$ denotes all neighbouring nodes of i without the node j. Once the algorithm has converged, the two-nodes beliefs that approximate the pair-marginals are calculated with

$$b_{ij}(x_i, x_j) = c \psi_{ij}(x_i, x_j) \phi_i(x_i) \phi_j(x_j) \times \prod_{k \in N(i) \setminus j} m_{ki}(x_i) \prod_{l \in N(j) \setminus j} m_{kj}(x_j), \qquad (4)$$

where c is a normalisation constant. The one-node beliefs for the approximation of the marginals $p_i(x_i)$ can be calculated with

$$b_i(x_i) = c \phi_i(x_i) \prod_{j \in N(i)} m_{ji}(x_i). \qquad (5)$$

Obviously, (5) is consistent with $b_i(x_i) = \sum_{x_j} b_{ij}(x_i, x_j)$, which also allows to calculate $b_i(x_i)$ from (4).

3 Superparamagnetic Clustering and Sequential Clustering

Superparamagnetic clustering

The superparamagnetic clustering algorithm is rooted in statistical physics, more precisely in *inhomogeneous Potts spin models*; these are extended Ising models for magnetic materials. Essentially, such models consist of interacting spins that are assigned to the sites of an irregular grid. For the clustering setting, the sites are given by the N data items to be clustered. Each spin variable x_i can take $q > 1$ values from a discrete set, i.e., $x_i \in \{1, ..., q\}$. The choice of q is largely arbitrary and is not connected to the number of occurring clusters. The spin sites are locally connected: each spin interacts with its k (not necessarily mutual) nearest neighbours. The neighbours are determined by means of the given distances or 'dissimilarities' d_{ij} between two items. SC can perform solely on the set of d_{ij}. The actual items need not be known. The coupling strength J_{ij} decreases with increasing d_{ij}. We choose

$$J_{ij} = J_{ji} = \frac{1}{\widehat{K}} e^{\frac{-d_{ij}^2}{2a^2}}, \qquad (6)$$

where $\widehat{K}$ is the average number of coupled neighbours (not necessarily equal to k) and a is the average distance between them. The chosen connectivity leads to a highly inhomogeneous network; this is a basic requirement for clustering and is in contrast to the normal situation of magnetic systems on regular grids. Each spin configuration is characterised by an energy value given by

$$H(\{x\}) = \sum_{(i,j)} J_{ij} (1 - \delta_{x_i x_j}), \qquad (7)$$

where the sum runs over all connections (ij). The expression (7) is referred to as *Potts spin Hamiltonian.* The system is considered in the formalism of the canonical ensemble. Thus, the probability for a certain spin configuration is given by the Boltzmann/Gibbs distribution

$$p(\{x\}) = \frac{1}{Z} e^{-H(\{x\})/T}, \qquad (8)$$

where the partition function $Z = Z(T)$ serves as a normalisation factor. The temperature T acts as a control parameter expressing the average energy of the system. As T is increased, the system typically undergoes a number of phase transitions: (**I**) For small T, the system is in the ferromagnetic phase where spins are likely to be aligned. (**II**) For an intermediary T-range, a superparamagnetic phase occurs. Strongly coupled spins tend to be aligned, whereas weakly coupled spins behave independently. Thus, clusters of aligned spins occur, reflecting groups of similar data points. A further increase of T generally leads to a cascade of these clusters into smaller clusters, so that a hierarchy of classes and subclasses is obtained. (**III**) For high T, the system enters the paramagnetic phase where any order disappears and only singleton clusters remain.

Among the data points, clusters can be identified for each T by means of the pair correlation criterion: Two points i and j belong to the same cluster, if the pair correlation

$$G_{ij} = \sum_{\{x\}} p(\{x\})\delta_{x_i x_j} = \sum_{x_i} p_{ij}(x_i, x_j = x_i) \quad (9)$$

exceeds a given threshold Θ ($\Theta = 0.7$ in this contribution). Clusters define transitive relations in the sense that if i and j as well as j and k belong to a cluster, all points belong to the very same cluster.

In practice, the sum (9) cannot be carried out for large sets. In the standard SC approach [1], a MCMC method (Swendsen-Wang) was proposed for an approximative calculation of (9). In the following, we combine the concept of PMRF and SC to replace MCMC by loopy BP.

SC translated into a PMRF

For practical reasons we formally amend the Hamiltonian (7) with an external field term, $H(\{x\}) = \sum_{(i,j)} J_{ij}(1 - \delta_{x_i x_j}) - \sum_i h_i(x_i)$, where $h_i(x_i) = 0$. The Boltzmann distribution (8) can then be factorised in the form

$$p(\{x\}) = \frac{1}{Z} \prod_{(ij)} e^{-E_{ij}(x_i,x_j)/T} \prod_i e^{h_i(x_i)/T}, \quad (10)$$

where $E_{ij}(x_i, x_j) = J_{ij}(1 - \delta_{x_i x_j})$. A comparison with (1) allows for a straightforward conversion of the Potts spin model into a PMRF. The factor functions read $\psi_{ij}(x_i, x_j) = e^{-E_{ij}(x_i,x_j)/T}$ and $\phi_i(x_i) = e^{h_i(x_i)/T}$.

Sequential clustering

For the purpose of data clustering, the superparamagnetic phase is the most relevant. In that phase, clusters of groups of similar data points appear. However, as T is increased, such clusters may break up into smaller units. Then the question arises: Which T provides the best clustering resolution, i.e., the most natural choice of clusters? However, this problem is ill-posed, since the clusters might be chosen from different T's. A hint on which clusters should be selected is given by their T-stability. The most natural cluster structures in the data set are stable over larger temperature ranges. In [4], we introduced an algorithm that successively extracts the most stable cluster(s) from a set and reclusters the single subsets (i.e., the cluster(s) and the residual set) with readjusted parameters $\widehat{K}$ and a in (6). The procedure stops in a branch if no stable (in terms of T-stability) substructures can be found anymore. This method does not only provide us with an intrinsic criterion for an automatic selection of clusters. In [4] it was also shown that this procedure is able to find natural substructures that would remain hidden due to density differences when only the whole set is clustered (for more details see [4]).

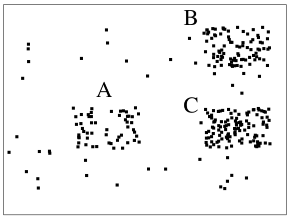

Figure 1: A simple two-dimensional toy system.

4 The Potential of Belief Propagation for Sequential Clustering

A toy system example

To gain a first estimate for the performance of BP in connection with sequential SC, we designed a simple two-dimensional toy system with 320 points (Fig. 1). The distribution contains three clusters (labelled by A,B,C) whose sizes are around $55, 90$ and 115 points. Some points cannot be clearly assigned to a cluster and form a background distribution.

In the following, we report on the comparison of BP with the standard MCMC algorithm (Swendsen-Wang). We used $k = 9$ and $q = 2$. The questions of interest are: **a)** Are the results achieved with BP as good as the results from the re-

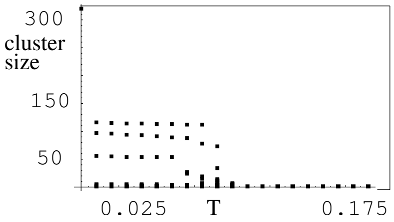

Figure 2: 'Naive' BP implementation applied to the toy set shown in Fig. 1. The size of occurring clusters is drawn as a function of T.

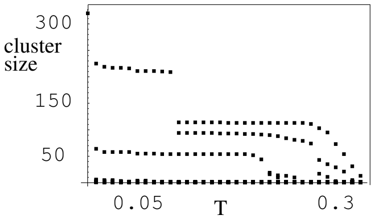

Figure 3: Results for BP with external field.

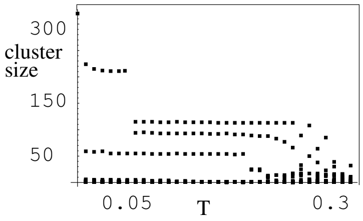

Figure 4: Results for MCMC.

liable MCMC? **b)** Is BP more efficient, i.e., quicker than MCMC? To judge a), it is sufficient for the sequential clustering procedure that size and T-stability of the stable clusters for BP and MCMC are very similar, but not necessarily identical.

Implementation issues: 1) A crucial step is the normalisation of the messages after each iteration, i.e., $\sum_{x_j} m_{ij}(x_j) = 1$. If one fails to do this, the messages quickly tend to zero. **2)** In Fig. 2 the results for a 'naive' BP implementation are displayed. On a first glance, the results seem reasonable; three clusters are clearly recognisable. However, these results only reflect the local potentials, i.e., $b_{ij}(x_i, x_j) = c\psi_{ij}(x_i, x_j)$, and can thus be achieved by a simple cut of all connections J_{ij} with $1/(1 + e^{-J_{ij}/T}) < \Theta$. This effect is due to the sustained symmetry of the messages for a uniform initialisation. For improvements, the symmetry has to be broken. In practice, we realised this by introducing a nonzero external field at one arbitrary site l in an arbitrary 'direction' $k: \ h_l(x_l) = \delta_{x_l k}$ (in contrast, one spin is totally clamped in [2]). This external bias leads to an asymmetry travelling through the whole (irreducible) network yielding improved results (Fig. 3), comparable to those of MCMC (Fig. 4).

Performance comparison: a) Fig. 3 and 4 are interpreted as follows: For $T = 0$ all points are in one cluster (ferromagnetic phase). This cluster immediately breaks up into two big units (one unit containing B and C, the other unit corresponding to A) and some singletons (background). After another transition, the three clusters A, B and C coexist and finally decay. The temperature of the decay depends on the compactness of a cluster, i.e., more compact clusters are more stable. The comparison of BP and MCMC yields almost identical results for the particular test set. The occurring differences are irrelevant for sequential clustering.

b) For sequential clustering, MCMC based on Swendsen-Wang is relatively efficient, i.e., 200 MC steps yield reliable results for the cluster detection. In the above example BP converges within 15 to about 300 iterations, depending on T. For higher T's, the algorithm slows down. The performance also depends on the site where the external field is placed. If it is placed within the background, the performance is worse. Our experiments so far have shown that our version of loopy BP does not necessarily outperform MCMC in terms of computational time. However, we hope to tap the full efficiency potential of BP in our further work.

5 CONCLUSIONS

In earlier work, we have shown how 'natural' clusterings can be determined without fixing the number of clusters a priori; the key idea is to sequentially extract clusters that persist over a large temperature range, and to recluster them separately. In this paper we exploited the potential of BP-based superparamagnetic clustering in connection with this sequential procedure. We reported on implementation and performance issues. In the near future, we will elaborate these issues in a more rigid manner. Furthermore, we plan to apply our approach to real-world tasks such as visual scene analysis and data analysis in combinatorial chemistry.

References

[1] M. Blatt, S. Wiseman, E. Domany, "Superparamagnetic clustering of data," *Phys. Rev. Lett.*, vol.76, pp.3251-3254, 1996.

[2] N. Shental, A. Zomet, T. Hertz, Y. Weiss "Pairwise Clustering and Graphical Models", Proceedings of NIPS 2003.

[3] J. S. Yedidia, W. T. Freeman,Y. Weiss, "Understanding Belief Propagation and its Generalizations", Exploring Artificial Intelligence in the New Millennium, ISBN 1558608117, Chap.8,pp.239-236 2003.

[4] T. Ott, A. Kern, A. Schuffenhauer, M. Popov, P. Acklin, E. Jacoby, R. Stoop, "Sequential superparamagnetic clustering for unbiased classification of high-dimensional chemical data," *J. Chem. Inf. Comput. Sci.*, vol. 44(4), pp. 1358-1364, 2004.