Discrete Combinatorial Optimisation in MAP Estimation

Tran The Truyen
thetruyen.tran@postgrad.curtin.edu.au

April 27, 2008

Contents

1 Problem Statement

Given N discrete variables $x_i \in \{1, 2, ..., S\}$ for $i = 1, 2, ..., N$. Let $\mathcal{E}$ be the set of pair of indices (i, j) where $i \neq j$. We want to find the set of optimal solutions as follows

$$(x_1^*, x_2^*, ..., x_N^*) = \arg \min_{x_1, x_2, \ldots, x_N} \left(\sum_{i \in [1,N]} g_i(x_i) + \sum_{(i,j) \in \mathcal{E}} f_{ij}(x_i, x_j) \right) \quad (1)$$

The RHS is often referred to as *energy*, a term borrowed from physical science. This equation is also known as *Maximum A Posteriori* (MAP) estimation, for the reason we will explain in Section 2.2.

2 Applications

2.1 Pattern Recognition

This problem has a wide application in pattern recognition (e.g. see [13, 5, 9, 12, 15, 3, 7, 10]). For example, given an image y of size $H \times W$. Thus $N = H \times$W is the number of pixels in the image. The set $\mathcal{E}$ includes all pairs of adjacent pixels in a grid. We want to give each pixel a label in the set $\{1, 2, ..., S\}$. Let x_i be a discrete variable that represents the label of the pixel y_i. The function $g_i(x_i)$ is now extended to include y_i as the input, i.e. $g_i(x_i, y_i)$ but this will not change the nature of Equation 1. The function $f_{ij}(x_i, x_j)$ is to encode the dependency between adjacent labels. An example of this dependency is

$$f_{ij}(x_i, x_j) = \delta(x_i, x_j)$$

where $\delta(x_i, x_j)$ is the Kronecker function, i.e. $\delta(x_i, x_j) = 1$ if $x_i = x_j$ and 0 otherwise. Sometimes, we may limit f_{ij} to the class of *metric* functions, i.e.

$$f_{ij}(x_i, x_j) \leq f_{ik}(x_i, x_k) + f_{kj}(x_k, x_j) \tag{2}$$

for any $k \neq i, j;\;\; k = 1, 2, ..., N$.

2.2 Markov Random Fields

Markov Random Fields are popular tool for modelling spatial stochastic processes (e.g. see [1, 8]). The joint variable $x = (x_1, x_2, ..., x_N)$ is modelled as a random variable with the following distribution

$$\Pr(x) = \frac{1}{Z} \prod_{(i,j) \in \mathcal{E}} \psi_{ij}(x_i, x_j) \tag{3}$$

where $Z = \sum_{x_1, x_2, ..., x_N} \prod_{(i,j)\in\mathcal{E}} \psi_{ij}(x_i, x_j)$ is the normalisation term, i.e. $\sum_x \Pr(x) = 1$. This is called the *prior* distribution. Often, the label x_i is known as the state that generates the observation y_i. The generation is modelled in a distribution $\Pr(y_i|x_i)$. The common question is that given y, we want to find the assignment of x that is most explained by y, i.e.

$$x^* = \arg\max_x \Pr(x|y)$$

$\Pr(x|y)$ is called the *posterior* distribution because it is computed *after* seeing y while the original $\Pr(x)$ is computed *before* seeing y.

By probability theory, we have the Bayes rule

$$\Pr(x|y) = \frac{1}{\Pr(y)} \Pr(y|x) \Pr(x)$$

so

$$x^* = \arg\max_x \Pr(y|x) \Pr(x) \tag{4}$$

since $\Pr(y)$ does not depend on x. The common assumption made in MRFs is that

$$\Pr(y|x) = \prod_i \Pr(y_i|x_i) \tag{5}$$

Substituting (3,5) into (4) yields

$$\begin{aligned} x^* &= \arg\max_x \prod_i \Pr(y_i|x_i) \prod_{(i,j)\in\mathcal{E}} \psi_{ij}(x_i,x_j) \\ &= \arg\min_x \left(\sum_i -\log \Pr(y_i|x_i) + \sum_{(i,j)\in\mathcal{E}} -\log \psi_{ij}(x_i,x_j) \right) \end{aligned}$$

where we have ignore Z because it does not depend on x. This clearly has the form of Equation 1.

3 State-of-the-Arts

There are number of solutions for this combinatorial problem (e.g. see [14] for an evaluation). The most well-known methods are:

- Iterated Conditional Mode [2]: this is a local greedy search method. Thus it is sensitive to initialisation and is often trapped in local minima.
- Simulated Annealing [6, 4]: this is guaranteed to find the optimal solution, but is often too slow in practice.
- Belief Propagation, also known as Max-Product algorithm: this is originated from Artificial Intelligence [11], often yields good solutions but may not converge.
- Graph-cuts [3]: this is currently the best family of algorithms for problems with metric functions $f_{ij}(x_i, x_j)$ (see Equation 2).

References

[1] Julian Besag. Spatial interaction and the statistical analysis of lattice systems (with discussions). *Journal of the Royal Statistical Society Series B*, 36:192–236, 1974.

[2] Julian Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society B*, 48(3):259–302, 1986.

[3] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(11):1222–1239, 2001.

[4] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 6(6):721–742, 1984.

[5] Jason K. Johnson, Dmitry Malioutov, and Alan S. Willsky. Lagrangian relaxation for MAP estimation in craphical models. In *45th Annual Allerton Conference on Communication, Control and Computing*, September 2007.

[6] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, May 1983.

[7] Vladimir Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28(10):1568–1583, Oct 2006.

[8] S.L. Lauritzen. *Graphical Models.* Oxford Science Publications, 1996.

[9] M. Leordeanu and M. Hebert. Efficient MAP approximation for dense energy functions. In *Proceedings of the 23rd International Conference on Machine learning (ICML)*, pages 545–552. ACM Press New York, NY, USA, 2006.

[10] Talya Meltzer, Chen Yanover, and Yair Weiss. Globally optimal solutions for energy minimization in stereo vision using reweighted belief propagation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 428–435, 2005.

[11] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Francisco, CA, 1988.

[12] P. Ravikumar and J. Lafferty. Quadratic programming relaxations for metric labeling and Markov random field MAP estimation. In *Proceedings of the 23rd international conference on Machine learning (ICML)*, pages 737–744. ACM Press New York, NY, USA, 2006.

[13] Jian Sun, Nan-Ning Zheng, and Heung-Yeung Shum. Stereo matching using belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 25(7):787–800, Jul 2003.

[14] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for Markov random fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, number 3952 in Lecture Notes in Computer Science, pages 16–29, 2006.

[15] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. Map estimation via agreement on (hyper)trees: Message-passing and linear-programming approaches. Technical Report UCB/CSD-03-1269, UC Berkeley CS Division, August 2003.