The Information Bottleneck Method[1]

Naftali Tishby and Fernando C. Pereira and William Bialek

Summarized by Yang, Zhi-rong on 2, March 2004

1 Introduction

A fundamental problem in formalizing our intuitive ideas about information is to provide a quantitative notion of "meaningful" or "relevant" information.

It is argued in this paper that information theory, in particular lossy source compression, provides a natural quantitative approach to the question of "relevant information." Specifically, the authors formulate a variational principle for the extraction or efficient representation of relevant information.

The standard analysis of lossy source compression is "rate distortion theory," which provides a tradeoff between the rate, or signal representation size, and the average distortion of the reconstructed signal. Rate distortion theory determines the level of inevitable expected distortion, D, given the desired information rate, R, in terms of the rate distortion function $R(D)$. The main problem with rate distortion theory is in the need to specify the distortion function first, which in turn determines the relevant features of the signal. Those features, however, are often not explicitly known and an arbitrary choice of the distortion function is in fact an arbitrary feature selection.

In this paper the authors formalized this intuitive idea using an information theoretic approach which extends elements of rate distortion theory. They also derived self consistent equations and an iterative algorithm for finding representations of the signal that capture its relevant structure, and prove convergence of this algorithm.

2 Review of rate distoration theorem

2.1 Mutual informatino as quantization rate

Let X denote the signal (message) space with a fixed probability measure $p(x)$, and let $\tilde{X}$ denote its quantized codebook or compressed representation. For each value $x \in X$ we seek a possibly stochastic mapping to a representative, or codeword in a codebook, $\tilde{x} \in \tilde{X}$, characterized by a conditional p.d.f. $p(\tilde{x}|x)$. The mapping $p(\tilde{x}|x)$ induces a soft partitioning of X in which each block is associated with the codebook elements $x \in X$, with probability $p(\tilde{x}|x)$.

The first factor that determines the quality of a quantization is the rate, or the average number of bits per message needed to specify an element in the codebook without confusion. This number per element of X is bounded from below by the mutual information

$$I(X;\tilde{X}) = \sum_{x\in X}\sum_{\tilde{x}\in\tilde{X}} p(x,\tilde{x})\log\left[\frac{p(\tilde{x}|x)}{p(\tilde{x})}\right]$$

since the average cardinality of the partitioning of X is given by the ratio of the volume of X to that of the mean partition, $2^{H(X)}/2^{H(X|\tilde{X})} = 2^{I(X;\tilde{X})}$, via the standard asymptotic arguments. Notice that this quantity is different from the entropy of the codebook, $H(\tilde{X})$, and this entropy is normally not the target to be minimized.

2.2 Problem to sovle in rate distortion theory

The partitioning of X induced by the mapping $p(\tilde{x}|x)$, has an expected distortion

$$\big\langle d(x,\tilde{x})\big\rangle_{p(x,\tilde{x})} = \sum_{x\in X}\sum_{\tilde{x}\in\tilde{X}} p(x,\tilde{x})d(x,\tilde{x})$$

There is a monotonic tradeoff between the rate of the quantization and the expected distortion: the larger the rate, the smaller is the achievable distortion.

The rate distortion theorem of Shannon and Kolmogorov characterizes this tradeoff through the rate distortion function, $R(D)$, defined as the minimal achievable rate under a given constraint on the expected distortion:

$$R(D) \equiv \min_{\{p(\tilde{x}|x):\big\langle d(x,\tilde{x})\big\rangle\leq D\}} I(X;\tilde{X})$$

or the variational form by attaching the Lagrange multiplier, β

$$\mathcal{F}\big[p(\tilde{x}|x)\big] = I(X;\tilde{X}) + \beta\big\langle d(x,\tilde{x})\big\rangle_{p(x,\tilde{x})}$$

2.3 Main result of Blahut-Arimoto algorithm

The optimal mapping in the rate distoration problem can be obtained by the BA algorithm in an iterative manner.

- Self consistent equations for iterations:

$$\begin{cases} p_{t+1}(\tilde{x}) = \sum_x p(x)p_t(\tilde{x}|x) \\ p_t(\tilde{x}|x) = \frac{p_t(\tilde{x})}{Z_t(x,\beta)}\exp(-\beta d(x,\tilde{x})) \end{cases}$$

3 Problem to solve in this paper

The problem of relevant quantization has to be addressed directly, by preserving the relevant information about another variable. The relevance variable, denoted here by Y, must not be independent from the original signal X, namely they have positive mutual information $I(X;Y)$. It is assumed here that we have access to the joint distribution $p(x,y)$, which is part of the setup of the problem.

The goal is to do quantization and capture as much of the information about Y as possible. The amount of information about Y in $\tilde{X}$ is given by $I(\tilde{X};Y) \leq I(X;Y)$

The quantization can be regarded as some kind of lossy compression, which cannot convey more information than the original data. As with rate and distortion, there is a tradeoff between compressing the representation and preserving meaningful information, and there is no single right solution for the tradeoff. The assignment to look for in this paper is the one that keeps a fixed amount of meaningful information about the relevant signal Y while minimizing the number of bits from the original signal X (maximizing the compression).[1] In effect the information that X provides about Y is passed through a "bottleneck" formed by the compact summaries in $\tilde{X}$.

The optimal assignment by minimizing the functional

$$\mathcal{L}[p(\tilde{x}|x] = I(\widetilde{X};X) - \beta I(\widetilde{X};Y)$$

where β is the Lagrange multiplier attached to the constrained meaningful information. information, while maintaining the normalization of the mapping $p(\tilde{x}|x)$ for every x. At $\beta = 0$ our quantization is the most sketchy possible–everything is assigned to a single point–while as $\beta \to \infty$ we are pushed toward arbitrarily detailed quantization. By varying the (only) parameter β one can explore the tradeoff between the preserved meaningful information and compression at various resolutions.

4 Main result of this paper

- Funtional to be minimized:

$$\mathcal{F}\big[p(\tilde{x}|x); p(\tilde{x}); p(y|\tilde{x})\big] = I(\widetilde{X};X) + \beta\big\langle D_{KL}\big[p(y|x)|p(y|\tilde{x})\big]\big\rangle_{p(x,\tilde{x})}$$

- Self consistent equations for iterations:

$$\begin{cases} p_t(\tilde{x}|x) = \frac{p_t(\tilde{x})}{Z_t(x,\beta)}\exp(-\beta d(x,\tilde{x})) \\ p_{t+1}(\tilde{x}) = \sum_x p(x)p_t(\tilde{x}|x) \\ p_{t+1}(y|\tilde{x}) = \sum_y p(y|x)p_t(x|\tilde{x}) \end{cases}$$

[1] It is completely equivalent to maximize the meaningful information for a fixed compression of the original variable.

- The structure of the solutions

  For every value of the Lagrange multiplier β there are corresponding values of the mutual information $I_X(X;\tilde{X})$, and $I_Y(\tilde{X};Y)$ for every choice of the cardinality of $\tilde{X}$. The variational principle implies that

$$\frac{\delta I_Y(\tilde{X};Y)}{\delta I_X(X;\tilde{X})} = \beta^{-1} > 0$$

  which suggests a deterministic annealing approach. By increasing the value of β one can move along convex curves in the "information plane" (I_X, I_Y). These curves, analogous to the rate-distortion curves, exists for every choice of the cardinality of $\tilde{X}$. The solutions of the self-consistent equations thus correspond to a family of such annealing curves, all start from the (trivial) point (0,0) in the information plane with infinite slope and are parameterized by the value of β. Interestingly, every two curves in this family separate (bifurcate) at some finite (critical) β, through a second-order phase transition. These transitions form an hierarchy of relevant quantizations for different cardinalities of $\tilde{X}$.

References

[1] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, 1999.