A study on difference of codelengths between MDL codes and Bayes codes on case different priors are assumed*

Masayuki GOTOH†, Toshiyasu MATSUSHIMA†, and Shigeichi HIRASAWA†
†Dept. of Industrial and Management Systems Engineering
Waseda University
Shinjuku-ku, Tokyo 169, Japan

Abstract — **We shall analyze quantitatively the the difference of codelengths between both codes for the hierarchical (nested) model class.**

I. Introduction

We shall discuss the two-step code based on the minimum description length principle (MDL code) proposed by J.Rissanen, which supposes implicitly the prior distribution. On the other hand, Bayes code (mixture code) [1] uses mixture of the probabilities in model class for coding function. If we can assume the same prior distribution, then the Bayes code is superior to the MDL code [3]. The properties of each codes have been studied independently. And we evaluated the difference of codelengths between both codes on condition that the identical prior distribution is assumed for both codes in the previous work [2].

In this paper, our previous results will be extended to the case that the different priors are assumed for both codes. We shall analyze quantitatively the the difference of codelengths between both codes for the hierarchical (nested) model class.

II. Preliminary Definition

Let $\mathcal{X}$ be the discrete source alphabet. Let $x^n = x_1 x_2 \cdots x_n$ be the data sequence with length n derived from the source, where $\forall i,\ x_i \in \mathcal{X}$. We shall discuss the ideal codelength measured by $-\log P(x^n)$.

Definition 1 (the hierarchical model class) *Let m be a discrete label of model in the discrete and finite model class $\mathcal{M}$, $m \in \mathcal{M}$. Each model has k_m-dimensional parameter $\theta^{k_m} \in \Theta^{k_m}$, then m specifies a parametric model class $\mathcal{H}^{k_m}$. Then the hierarchical model class $\mathcal{H}$ is defined by $\mathcal{H} = \cup_m \mathcal{H}^{k_m}$, where the nested structure $\mathcal{H}^{k_{m1}} \subset \mathcal{H}^{k_{m2}} \subset \cdots$ is satisfied for $m1, m2, \cdots \in \mathcal{M}$ and $k_{m1} < k_{m2} < \cdots$.* □

The data sequence x^n is derived from $P(x^n|m^*, \theta^{k^*_{m^*}})$, where parameter with $*$ shows those of the true. The main conditions required here are that the distribution of the maximum likelihood estimator and the Bayesian posterior probability of parameter have asymptotic normality.

Definition 2 (MDL codes and Bayes codes) *The two type MDL codes can be defined. At first, we define the MDL code quantizing parameter space and selecting both of quantized parameter and discrete label. The codelength of this type of the MDL code, $L^{m,\bar{\theta}^{k_m}}_{MDL}(x^n)$, is given by*

$$L^{m,\bar{\theta}^{k_m}}_{MDL}(x^n) = \min_{m,\bar{\theta}^{k_m}} \Big\{ -\log P(x^n|m,\bar{\theta}^{k_m}) - \log \frac{f_M(\bar{\theta}^{k_m}|m)}{\sqrt{n^k}\sqrt{\det(\bar{\theta}^{k_m}|m)}} - \log P_M(m) \Big\}. \quad (1)$$

Here $f_M(\cdot|m)$ and $P_M(\cdot)$ is prior density and prior probability for the MDL code, respectively.

Secondly, we define the MDL code using mixture for the parameter and selecting only a discrete label m.

$$L^{m,\theta^{k_m}}_{MDL}(x^n) = \min_m \Big\{ -\log q^{\theta^{k_m}}(x^n|m) - \log P_M(m) \Big\}, \quad (2)$$

$$-\log q^{\theta^{k_m}}(x^n|m) = -\log \int_{\theta^{k_m}} P(x^n|m,\theta^{k_m}) f_M(\theta^{k_m}|m) d\theta^{k_m}. \quad (3)$$

On the other hand, the codelength of the Bayes code is given by

$$L^{m,\theta^{k_m}}_{Bayes}(x^n) = -\log \sum_m \int_{\theta^{k_m}} P(x^n|m,\theta^{k_m}) f_B(\theta^{k_m}|m) P_B(m) d\theta^{k_m}. \quad (4)$$

Here $f_B(\cdot|m)$ and $P_B(\cdot)$ are prior density and prior probability for the Bayes code, respectively, and the integral is calculated through the parameter set Θ^{k_m}. □

III. Main Results

In this section, we shall analyze the difference of the codelengths between the MDL code and the Bayes code on condition the diffrent priors are assumed for both codes.

Theorem 1 *On suitable conditions, the following inequations are satisfied for sufficient large n.*

If $P_M(m^*) > P_B(m^*)$, then

$$L^{m,\theta^{k_m}}_{MDL}(x^n) < L^{m,\theta^{k_m}}_{Bayes}(x^n), \quad (5)$$

Else if $P_M(m^*) \le P_B(m^*)$, then

$$L^{m,\theta^{k_m}}_{MDL}(x^n) > L^{m,\theta^{k_m}}_{Bayes}(x^n). \quad (6)$$

□

This result shows that an increment of codelength caused by selecting a model m converges to 0. Therefore the code assuming advantageous prior is superior to the other.

Theorem 2 *On suitable condition, if*

$$\log \frac{f_B(\theta^{k^*_{m^*}}|m^*) P_B(m^*)}{f_M(\theta^{k^*_{m^*}}|m^*) P_M(m^*)} > \log R(\sigma) + o(1), \quad (7)$$

*is satisfied,** the following relation is asymptotically satisfied.*

$$L^{m,\bar{\theta}^{k_m}}_{MDL}(x^n) > L^{m,\theta^{k_m}}_{Bayes}(x^n). \quad (8)$$

Here, $R(\sigma)$ is some positive constant. □

In this case, the Bayes code is superior to the MDL code assuming advantageous prior if this advantage of the MDL code is essentially not too strong.

References

[1] B.S.Clarke and A.R.Barron : "Jeffreys' Prior is Asymptotically Least Favorable under Entropy Risk", JSPI 41, pp.37-60, (1994)

[2] M.Gotoh, T.Matsushima, and S.Hirasawa: "An Analysis on Difference of Codelengths between Codes Based on MDL Principle and Bayes Codes", Submitted to IEICE Trans. Fundamentals

[3] J.Rissanen: "Stochastic Complexity", J. Roy, Statist., Soc.B, Vol.49, pp.223-265, (1987)

*This work was supported in part by the Ministry of Education under Grant-Aids 07558168 for Science Research and by Waseda University under Grant 96A-259 for Special Research Projects.

**Even if $f_B(\theta^{k^*_{m^*}}|m^*)P_B(m^*) < f_M(\theta^{k^*_{m^*}}|m^*)P_M(m^*)$ is satisfied, (8) is satisfied when (7) is also satisfied.