THE MDL PRINCIPLE FOR ARBITRARY DATA: EITHER DISCRETE OR CONTINUOUS OR NONE OF THEM

Joe Suzuki

Department of Mathematics, Graduate School of Science, Osaka University,
1-1 Machikaneyama-cho Toyonaka, Osaka 560-0043, Japan

ABSTRACT

Consider universal data compression: the length $l(x^n)$ of sequence $x^n \in A^n$ with finite alphabet A and length n satisfies Kraft's inequality over A^n, and $-\frac{1}{n}\log\frac{P^n(x^n)}{Q^n(x^n)}$ almost surely converges to zero as n grows for the $Q^n(x^n) = 2^{-l(x^n)}$ and any stationary ergodic source P. In this paper, we say such a Q is a universal Bayesian measure. We generalize the notion to the sources in which the random variables may be either discrete, continuous, or none of them. The previous work by Ryabko was applied only to the case that a density function exists, and an additional constraint on P was required. The universal Bayesian measure constructed in this paper has many applications to infer relation among random variables, and extends the application area of the minimum description length principle.

1. SUMMARY

Suppose we wish to know if discrete random variables X, Y are independent ($X \perp\!\!\!\perp Y$) given n pairs of examples $\{(x_i, y_i)\}_{i=1}^n$ emitted by (X, Y). If the probabilities of $x^n = (x_1, \cdots, x_n)$, $y^n = (y_1, \cdots, y_n)$, and (x^n, y^n) are expressed by $P_X^n(x^n|\theta_X)$, $P_Y^n(y^n|\theta_Y)$, and $P_{XY}^n(x^n, y^n|\theta_{XY})$, respectively, using unknown parameters $\theta_X, \theta_Y, \theta_{XY}$, one way to deal with this problem is to decide $X \perp\!\!\!\perp Y$ if and only if

$$pQ_X^n(x^n)Q_Y^n(y^n) \geq (1-p)Q_{XY}^n(x^n, y^n)\ ,$$

where p is the prior probability of $X \perp\!\!\!\perp Y$, and the three values are defined by

$$\begin{aligned} Q_X^n(x^n) &:= \int P^n(x^n|\theta_X)w_X(\theta_X)d\theta_X\ , \\ Q_Y^n(y^n) &:= \int P^n(y^n|\theta_Y)w_Y(\theta_Y)d\theta_Y\ , \\ Q_{XY}^n(x^n, y^n) &:= \int P^n(x^n, y^n|\theta_{XY})w_{XY}(\theta_{XY})d\theta_{XY} \quad (1) \end{aligned}$$

using weights w_X, w_Y, w_{XY} over the parameters θ_X, θ_Y, θ_{XY}, respectively.

To this end, let A be the finite set in which X takes values. There are many options of Q_X such that

$$\sum_{x^n \in A^n} Q_X^n(x^n) \leq 1\ . \quad (2)$$

For example[1], $Q_X^n(x^n) = |A|^{-n}$ for $x^n \in A^n$ satisfies the condition. However, such a Q_X cannot be an alternative of P for large n because Q_X^n does not converges to P^n in any sense. On the other hand, if we choose $w_X(\theta_X) \propto \prod_{x\in A}\theta_x^{-a[x]}$ with constants $(a[x] = \frac{1}{2})_{x\in A}$ (Krichevsky-Trofimov [3]), then the quantity $-\frac{1}{n}\log Q_X^n(x^n)$ almost surely converges to its entropy $H(\theta_X)$ for any independent and identically distributed (i.i.d) source $P^n(x^n|\theta_X) = \prod_{x\in A}\theta_x^{-c[x]}$ with parameters $\theta = (\theta_x)_{x\in A}$ and frequencies $(c[x])_{x\in A}$ in $x^n \in A^n$ [6]. Furthermore, the Shannon-McMillian-Breiman theorem [2] states that

$$-\frac{1}{n}\log P^n(x^n|\theta_X) \to H(\theta_X)$$

almost surely for any stationary ergodic source θ_X, so that almost surely

$$\frac{1}{n}\log\frac{P_X^n(x^n)}{Q_X^n(x^n)} \to 0 \quad (3)$$

if we write $P^n(x^n|\theta_X)$ by $P_X^n(x^n)$. In this paper, we say such a Q_X satisfying (2)(3) to be a *universal Bayesian measure* associated with finite set A. From the above discussion, we can say that a universal Bayesian measure exists for finite sources.

However, what if X, Y are arbitrary without assuming they are discrete? Recently, for random variable X such that its density function f_X exists, Boris Ryabko [5] proved that there exists g_X such that

$$\int_{x^n\in\mathbb{R}^n} g_X^n(x^n) \leq 1$$

and

$$\frac{1}{n}\log\frac{f_X^n(x^n)}{g_X^n(x^n)} \to 0\ . \quad (4)$$

for any f_X satisfying a condition. The estimation is based on a specific sequence of histograms: let $A_0 = \{A\}$, and A_{j+1} is a refeinment of A_j, where A is the range of X; for each $j = 1, 2, \cdots$, we estimate the density function $f_j^n(x^n)$ based on each histogram A_j; and if we obtain $g_j^n(x^n)$ as the estimation for $j = 1, 2, \cdots$, then we obtain the final estimation $g^n(x^n) = \sum_j w_j g_j^n(x^n)$, where $w_j > 0$ and $\sum_j w_j = 1$. But the estimation depends on the specific $\{A_j\}$ and requires $D(f_X||f_j) \to 0$

[1] $|A|$ denotes the cardinality of set A.

as $j \to \infty$. Therefore, sufficient prior information is required.

In addition, in order to decide whether $X \perp\!\!\!\perp Y$ or not is made, we need to construct Bayesian measures Q_{XY} and g_{XY} for two variables X, Y extending Q_X and g_X for one variable X.

We admire Ryabko's original work [5], and admit that the basic idea was already there. However, we need to seek further generalizations for practical development of the theory. In this paper, we

1. remove the constraint that X should be either discrete or continuous to obtain a general form of universality containing (3)(4) as special cases;
2. remove the condition that Ryabko [5] posed; and
3. construct universal measures for more than one variables,

so that we establish that a universal Bayesian measure unconditionally exists for any stationary ergodic random variable which may be either discrete, continuous, or none of them. Once we can deal with universal Bayesian measures for more than one random variables, we can infer relation among them from given examples.

For simplicity, in this paper, we assume that the underlying source is i.i.d. although the discussion will hold for stationary ergodic sources.

Let us state the main results without proof. Let $\mathcal{B}$ be the entire Borel sets of $\mathbb{R}$. Let $\mu_X(D) := P(X \in D)$ for $D \in \mathcal{B}$, and η_X a σ-finite measure, i.e. there exists $\{A_j\}$ such that $A_j \in \mathcal{F}, \cup_j A_j = \Omega$ and $\eta_X(A_i) < \infty$ for measure space $(\Omega, \mathcal{F})$, and we assume[2] $\mu_X \ll \eta_X$, i.e., $\eta_X(D) = 0 \Longrightarrow \mu_X(D) = 0$ for any $D \in \mathcal{B}$.

Theorem 1 There exists a $\nu_X^n \ll \eta_X^n$ such that $\nu_X^n(A^n) \leq 1$ and with probability one as $n \to \infty$

$$\frac{1}{n} \log \frac{d\mu_X^n}{d\nu_X^n}(x^n) \to 0 \tag{5}$$

We notice that the Radon-Nikodym derivative is expressed by the ratio

$$\frac{d\mu_X^n}{d\nu_X^n}(x^n) = \frac{d\mu_X^n}{d\eta_X^n}(x^n) / \frac{d\nu_X^n}{d\eta_X^n}(x^n)$$

and that the density function $\frac{d\mu_X^n}{d\eta_X^n}(x^n)$ in the generalized sense is estimated by constructing the quanity $\frac{d\nu_X^n}{d\eta_X^n}(x^n)$. If η_X is the Lebesgue measure of $\mathbb{R}$, then the result reduces to Ryabko's result $f_X = \frac{d\mu_X^n}{d\eta_X^n}(x^n)$ and $g_X = \frac{d\nu_X^n}{d\eta_X^n}(x^n)$. However, we have successfully removed the condition
$D(f_X || f_j) \to 0$ as $j \to \infty$.

In a similar way, let $\mu_Y(D) := P(Y \in D)$ for $D \in \mathcal{B}$, and η_Y a σ-finite measure, and we assume $\mu_X \ll \eta_X$. Furthermore, Let[3] $\mu_{XY}(D \times D') := P(X \in D, Y \in D')$ for $D, D' \in \mathcal{B}$, and η_{XY} the product measure of η_X, η_Y. Let B be the range of Y.

[2] We read that μ_X is absolutely continuous w.r.t. η_X.

[3] $D \times D'$ denotes the Cartesian product of sets D, D'.

Theorem 2 There exists a $\nu_{XY}^n \ll \eta_{XY}^n$ such that $\nu_{XY}^n(A^n \times B^n) \leq 1$ and with probability one as $n \to \infty$

$$\frac{1}{n} \log \frac{d\mu_{XY}^n}{d\nu_{XY}^n}(x^n, y^n) \to 0 \tag{6}$$

The results in this paper are rather theoretical but contain many applications such as

1. Bayesian network structure learning [7, 8],
2. a variant of the Chow-Liu algorithm learning a forest given examples [7, 9] .

In fact, in any database, both discrete and continuous fields are present. Then, we need to find dependency among those attributes. However, the existing results only dealt with either only discrete data or only continuous data. This paper deals with the most general and realistic cases.

For contributions to statistics, constructing such a universal Bayesian measure means establishing a general form of Bayesian Information Criteria (BIC). Suppose we have a countable number of models $m = 1, 2, \cdots$ each of which expresses a relation among random variables. If we construct a universal Bayesian measure $q(x^n|m)$ w.r.t. model m given data x^n, then we can select m such that $-\log p(m) - \log q(x^n|m)$ is minimized, where $p(m)$ is the prior probability of model m. In fact, the measure applies to all the cases that BIC/MDL applied thus far.

2. REFERENCES

[1] P. Billingsley. *Probability & Measure* (1995): (3rd ed.). New York : Wiley.

[2] T. M. Cover and J. A. Thomas. *Elements of Information Theory* (1995): (2nd ed.). New York : Wiley.

[3] R.E. Krichevsky and V.K. Trofimov, "The Performance of Universal Encoding", *IEEE Trans. Inform. Theory* 27(2): 199-207 (1981).

[4] J.Rissanen, "Modeling by shortest data description". *Automatica* 14: 465-471 (1978).

[5] B. Ryabko, "Compression-Based Methods for Nonparametric Prediction and Estimation of Some Characteristics of Time Series." *IEEE Trans. on Inform. Theory*, 55(9):4309-4315 (2009).

[6] B. Ryabko, "Prediction of random sequences and universal coding", *Problems Inform. Transmission* 24 (1988), no. 2, 87–96. Russian: *Problemy Peredachi Informatsii* 24 (1988), no. 2,3–14

[7] J. Suzuki, "A Construction of Bayesian Networks from Databases on an MDL Principle", *The Ninth Conference on Uncertainty in Artificial Intelligence*, Washington D. C., pages 266-273, 7 (1993).

[8] J. Suzuki, "The Universal Measure for General Sources and its Application to MDL/Bayesian Criteria", *Data Compression Conference* 2011, Snowbird, Utah (2011).

[9] J. Suzuki, "The Bayesian Chow-Liu Algorithms ", pages 315-322, Proceedings of the 6th workshop on Probabilistic Graphical Models, Granada, Spain. (2012)