Method Description

General Information

Type of Entry	Student
First Name	Roman
Last Name	Sirotin
Country	Russia
Type of Affiliation	University
Affiliation	Siberian State University of Telecommunications and Information Sciences

Information about the method utilized

Name of Method	R-method
Type of Method	Statistical
Short Description (up to 200 words)	R-method was proposed by Ryabko. It is based on the discovery of deep theoretical connection between data compression, time series forecasting, and estimating made by Rissanen.

R-method Extended Description

Contents

R-method was proposed in Ryabko [2008]. It is based on the discovery of deep theoretical connection between data compression, time series forecasting, and estimating made in Rissanen [1984].

1 R-method for discrete series

Let's consider a source generating sequences $x_1 x_2 \ldots$ of letters from some discrete alphabet A. Let the source generate a message $x_1 \ldots x_t$, $x_i \in A$, and next letter x_{t+1} needs to be predicted. The best possible point forecast would be $x_{t+1} \in A$ such that $\forall a \in A$: $p(a|x_1 \ldots x_t) \leq p(x_{t+1}|x_1 \ldots x_t)$.

Any γ that is an estimate of density p is said to be a predictor. Any predictor γ defines a measure by the following equation

$$\gamma(x_1 \ldots x_t) = \prod_{i=1}^{t} \gamma(x_i|x_1 \ldots x_{i-1}),$$

and vice versa (any measure or estimate of a measure) defines a predictor:

$$\gamma(x_i|x_1 \ldots x_{i-1}) = \frac{\gamma(x_1 \ldots x_{i-1} x_i)}{\gamma(x_1 \ldots x_{i-1})}.$$

It is shown in Krichevsky [1968] (see also Krichevsky [1993]) that Krichevsky predictor given by

$$K_0(a|x_1 \ldots x_t) = (\nu_{x_1 \ldots x_t}(a) + 1/2)/(t + |A|/2),$$

has the minimal error in some sense for the class of i.i.d sources.

Here $\nu_{x_1 \ldots x_t}(a)$ is the count of letter a occurrring in the word $x_1 \ldots x_t$.

This predictor can be extended on general Markov processes. The trick is to view a m-order Markov source with alphabet A as an i.i.d source with alphabet $|A|^m$. In that case Krichevsky predictor is given by

$$K_m(x_1 \ldots x_t) = \begin{cases} \frac{1}{|A|^t}, & t \leq m, \\ \frac{1}{|A|^m} \prod_{v \in A^m} \frac{\prod_{a \in A}((\Gamma(\nu_x(va)+1/2)/\Gamma(1/2)))}{(\Gamma(\overline{\nu}_x(v)+|A|/2)/\Gamma(|A|/2))}, & t > m \end{cases},$$

where $\overline{\nu}(v) = \sum_{a \in A} \nu_x(va)$, $x = x_1 \ldots x_t$.

Let us define the measure R, which is a consistent estimator of probabilities for the class of all stationary ergodic processes with a finite alphabet. First we define a probability distribution $\{\omega = \omega_0, \omega_1, \ldots\}$ by

$$\omega_i = 1/\log_2(i+2) - 1/\log_2(i+3),\ i \in \{0,\ 1,\ \ldots\}.$$

The measure R is defined as follows:

$$R(x_1 \dots x_t) = \sum_{i=0}^{\infty} \omega_i K_i(x_1 \dots x_t).$$

2 R-method for real-valued series

Let's consider a source generating values $x_1, x_2, \dots$ from interval $\Omega = (a, b) \subset \mathbb{R}$. And let $\{\Pi_n\}, n > 1$, be an increasing sequence of finite partitions of Ω, such that $\Pi_k = \{(a, a + h], (a+2h], \dots (a+(k-1)h, b), h = \frac{b-a}{k}\}$. Let $x^{[k]}$ denote the element of Π_k that contains the point x.

An estimation of denstiy of the real-valued source is then defined as

$$r_U(x_1 \dots x_t) = \sum_{i=0}^{\infty} \omega_i R(x_1^{[i]} \dots x_t^{[i]}) / M_t(x_1^{[i]} \dots x_t^{[i]}),$$

where $M_t(x_1^{[i]} \dots x_t^{[i]}) = (1/i)^t$.

3 Forecasts for M4

Equations for r_U and R contain infinite sums of decreasing terms. To produce any forecasts it is required to limit these sums to a finite number of addendums. We limit sum for r_U to $n = \lfloor \log_2(t) \rfloor + 5$, where t is a length of a series; and we limit sum for R to $m = 5$. Лысяк, Рябко [2016]

We first transform all series using Box-Cox transform. Then we apply differencing zero, one or two times depending on the results of Kwiatkowski–Phillips–Schmidt–Shin unit root test (but no more than two times). After that conditional distribution on Π_n is acquired using the following equation:

$$r_U(a_{mid} | x_1 \dots x_t) = r_U(x_1 \dots x_t a_{mid}) / r_U(x_1 \dots x_t), a \in \Pi_n,$$

where a_{mid} is a midpoint of $a \in \Pi_n$. $a \in \Pi_n$ with the highest conditional probability is then chosen and the point forecast is determined by calculating an average value of all the points from $\{x_1, x_2, \dots x_t\}$ that fall into $a^{[n]}$. If there are no such points a midpoint of $a^[n]$ is chosen as a forecast.

At the final step differencing and Box-Cox transform are reversed. The resulting point is added as the next forecasted value. The procedure is repeated until the multi-step forecast contains required number of points. After differencing is reversed the series may contain negative values, which makes reversing Box-Cox transform impossible. If such situation occurs the procedure is restarted without applying the Box-Cox transform on the first step.

References

1. Лысяк А., Рябко Б. Прогнозирование временных рядов на основе методов сжания данных // Проблемы передачи информации. 2016. Т. 52. № 1. С. 64–72.

2. Krichevsky R. A relation between the plausibility of information about a source and encoding redundancy // Problems Inform. Transmission. 1968. Т. 4. № 3. С. 48–57.

3. Krichevsky R. Universal Compression and Retrival.: Kluver Academic Publishers, 1993.

4. Rissanen J. Universal Coding, Information, Prediction, and Estimation // IEEE Transactions of Information Theory. 1984. Т. 30. № 4. С. 629–636.

5. Ryabko B. Compresion-based methods for nonparametric on-line prediction, regression, classification and density estimation of time series // 2008.