CHANGE DETECTION, HYPOTHESIS TESTING, AND DATA COMPRESSION

Kenji Yamanishi[1], *Ei-ichi Sakurai*[2], *Hiroki Kanazawa*[1]

[1]Graduate School of Information Science and Engineering, The University of Tokyo,
yamanishi@mist.i.u-tokyo.ac.jp, hiroki_kanazawa@mist.i.u-tokyo.ac.jp
[2] National Institute of Advanced Industrial Science and Technology, e.sakurai@aist.go.jp

ABSTRACT

We are concerned with the issue of detecting changes of statistical models when they change over time. We introduce the dynamic model selection (DMS) algorithm for learning model sequences on the basis of the minimum description length (MDL) principle. We first analyze it from the view of hypothesis testing. We evaluate error probabilities for testing the occurrences of change-points and relate them to the model transition estimators and the distance between the models to be distinguished. We then apply the DMS algorithm into data compression via piecewise stationary memoryless sources (PSMS's). We give a method for discretizing the parameter space to obtain an optimal data compression bound. From the both views of hypothesis testing and data compression, we argue how to discretize the parameter space in order to obtain ideal performance. It yields a new view of distinguishability of probabilistic models from the standpoint of change-detection.

1. INTRODUCTION

We are concerned with the issue of detecting changes of probabilistic models from a non-stationary data sequence. Dynamic model selection, which we abbreviate as DMS, has been proposed in [14],[13](see also [3]) in order to address this issue. DMS algorithms have been designed on the basis of the minimum description length (MDL) principle ([8]). I.e., they output a model sequence so that the sum of the code-length for a data sequence plus that for a model sequence is minimum. DMS is related to works by van Erven et.al.[2] on switching distributions, those by Shamir and Merhav [10], Willems [11],Willems and Casadei [12] on data compression for piecewise stationary memoryless sources (PSMSs).

In this paper we first analyze DMS from the view of hypothesis testing. We apply DMS to the issue of testing whether a change-point of statistical models exists or not, and evaluate it in terms of Type 1 and 2 error probabilities, which depend on how to estimate model transitions. We investigate them for the three types of methods for estimating model transition probabilities: Shamir and Merhav's method [10], Krichevsky and Trofimov's one [6], and Willem's one [11],[12].

We then apply DMS to data compression. We derive upper bounds on the total code-length for the three methods for estimating model transitions. We also apply DMS to learning piecewise stationary memoryless sources (PSMSs[9]) and analyze it from the view of data compression. According to [4], we give a method for discretizing the parameter space in order to get an optimal code-length bound. From the both views of hypothesis testing and data compression, we argue how to discretize the parameter space to obtain ideal performance. This yields a new insight into distinguishability([1],[8]) of probabilistic models from the view of change-detection as well as data compression.

2. DYNAMIC MODEL SELECTION

Following .[14],[13] we introduce a framework for DMS. Let $\mathcal{X}$ be a domain, which may be either continuous or discrete. Let x take a value in $\mathcal{X}$. Let $\mathcal{M}$ be a class of models, each of which is specified by a discrete parameter and is properly ordered. For example, we may consider the case where $M \in \mathcal{M}$ is a dimension of real-valued parameters. We denote $x_1 \ldots x_{t-1}$ as x^{t-1}. Let $P(X^n|M)$ be a probability distribution specified by a model M. For each M, for each t, we define a predictive distribution of X given x_a^b by $P(X|x^{t-1} : M) = P(X \cdot x^{t-1}|M)/P(x^{t-1}|M)$.

We suppose that a model switches to neighboring ones with some probabilities at each time. According to [14], we consider model transition probability distributions:

Definition 1 Let M range over $\{1, \ldots, \bar{M}\}$. Let α be a 1-dimensional parameter. Assuming that a model transits to neighbouring ones only, we define the *model transition probability distribution* as: $P(M_1|\emptyset : \alpha) = 1/\bar{M}$,

$$P(M_t|M^{t-1} : \alpha) = \begin{cases} 1-\alpha & \text{if } M_t = M_{t-1}, M_t \neq 1 \text{ or } \bar{M}, \\ 1-\alpha/2 & \text{if } M_t = M_{t-1}, M_t = 1 \text{ or } \bar{M}, \\ \alpha/2 & \text{if } |M_t - M_{t-1}| = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Here are three methods for estimating α.

Definition 2 *Shamir and Merhav's (SM) estimator* $\hat{\alpha}$ is defined as follows[10]: For $\epsilon > 0$,

$$\hat{\alpha}(M^t) = \frac{\pi(t - t_c + 1)}{Z_\infty - Z_{t-t_c}}, \tag{1}$$

where t_c is the latest change point before t and $\pi(t) = \frac{1}{t^{1+\epsilon}}$, $Z_n = \sum_{j=1}^{n} \pi(j)$, $Z_\infty = \sum_{j=1}^{\infty} \pi(j)$. *Krichevsky and Trofimov's (KT) estimator* $\hat{\alpha}$ is defined as follows[6]:

$$\hat{\alpha}(M^t) = (n(M^t) + 1/2)/t, \tag{2}$$

where $n(M^t)$ is the number of model changes in M^t. *Willem's (W) estimator* $\hat{\alpha}$ is defined as follows[11]:

$$\hat{\alpha}(M^t) = 1/(2(t - t_c)), \tag{3}$$

where t_c is the latest change-point before t.

KT estimator is calculated using all the past data, while SM and W estimators are calculated using the data starting from the latest change-point.

We denote $P(M_t|M^{t-1} : \hat{\alpha}(M^{t-1}))$ as $\hat{P}_t(M_t|M^{t-1})$. Below we give a criterion for selecting an optimal sequence on the basis of the MDL principle.

Definition 3 [14] Given $x^n = x_1 \dots x_n$, we define the *DMS criterion* for $M^n = M_1 \dots M_n$ by:

$$\begin{aligned}\ell(x^n : M^n) &= \sum_{t=1}^{n} \left(-\log P(x_t|x^{t-1} : M_t)\right) \\ &\quad + \sum_{t=1}^{n} \left(-\log \hat{P}_t(M_t|M_{t-1})\right). \end{aligned} \tag{4}$$

The first term is the total predictive code-length for x^n relative to M^n while the second term is the total predictive code-length for k^n. Hence the optimal sequence is obtained as the one which minimizes the total code-length. It leads to the DMS algorithm as follows:

Definition 4 [14] *The DMS algorithm*, denoted as DMS, is an algorithm that takes as input x^n and outputs $\hat{M}^n$ s.t.

$$\hat{M}^n = \arg\min_{M^n} \ell(x^n : M^n). \tag{5}$$

An algorithm that computes $\hat{M}^n$ as in (5) using the dynamic programming has been proposed ([14]).

3. HYPOTHESIS TESTING WITH DMS

We simplify the problem of DMS so that there are only two models; M_1 and M_2. We are then concerned with the issue of testing whether a model has changed or not. Below we assume that the model is either M_1 or M_2, the initial model is M_1, and there exists only one change-point in a model sequence. The problem is to detect when the model has changed. We give the following specific form of DMS in order to solve this issue.

Definition 5 *DMS as a change-point detector* is an algorithm that takes as input x^n and outputs the least time index t_c such that

$$\ell(x^n : M_1^n) \geq \ell(x^n : M^n(t_c)), \tag{6}$$

where $M^n(t_c) \stackrel{\text{def}}{=} \overbrace{M_1 \dots M_1}^{t_c} \overbrace{M_2 \dots M_2}^{n-t_c}$ and $M_1^n \stackrel{\text{def}}{=} M_1 \dots M_1$.

We reduce the change-detection problem to the hypothesis testing as follows: Let t^* be a true change-point. Consider the following two hypotheses: H_0 and H_1:

$$\begin{aligned} H_0 &: \quad M_1 \quad \text{for } x_1^n = x^n = x_1 \cdots x_n, \\ H_1 &: \quad \begin{cases} M_1 & \text{for } x_1^{t^*} = x_1 \cdots x_{t^*}, \\ M_2 & \text{for } x_{t^*+1}^n = x_{t^*+1} \cdots x_n. \end{cases} \end{aligned}$$

Set $P(x_{t^*+1}^n|x^{t^*} : M_1) \stackrel{\text{def}}{=} \prod_{j=t^*+1}^n P(x_j|x^{j-1} : M_1)$, and $P(x_{t^*+1}^n|x^{t^*} : M_2) \stackrel{\text{def}}{=} \prod_{j=t^*+1}^n P(x_j|x^{j-1} : M_2)$.Then DMS as a change-point detector works as a hypothesis testing algorithm such that H_0 is accepted if

$$\begin{aligned} &\sum_{t=t^*+1}^{n} \left(-\log P(x_t|x^{t-1} : M_1)\right) \\ &- \sum_{t=t^*+1}^{n} \left(-\log P(x_t|x^{t-1} : M_2)\right) < f(n, t^*), \end{aligned} \tag{7}$$

where

$$f(n, t^*) \stackrel{\text{def}}{=} \ell(M^n(t^*)) - \ell(M_1^n), \tag{8}$$

and

$$\ell(M_1^n) \stackrel{\text{def}}{=} \sum_{t=1}^{n} (-\log \hat{P}_t(M_1|M_1)),$$

$$\begin{aligned} \ell(M^n(t^*)) &\stackrel{\text{def}}{=} \sum_{t=1}^{t^*-1} (-\log \hat{P}_t(M_1|M_1)) + (-\log \hat{P}_{t^*}(M_2|M_1)) \\ &\quad + \sum_{t=t^*+1}^{n} (-\log \hat{P}_t(M_2|M_2)). \end{aligned}$$

Otherwise H_1 is accepted.

We define as measures of performance of a change-point detector Type 1 and 2 error probabilities as follows:

Definition 6 For given the length of data sequence n, the change-point time t^*, we define *Type 1 error probability* for DMS as a change-point detector by:

$$Prob\left[x_{t^*+1}^n \sim P(X^n|M_1) \text{ and Eq.(7) doesn't hold}\right],$$

and *Type 2 error probability* for DMS at delay $h = n - t^*$ by:

$$Prob\left[x_{t^*+1}^n \sim P(X_{t^*+1}^n|x^{t^*} : M_2) \text{ and Eq.(7) holds}\right].$$

Type 1 error probability is the probability that the model change has not yet occurred until time n but the change is incorrectly reported at time t^*. Type 2 error probability is the probability that the model change has already occurred at time t^*, but it is overlooked until time n where $h = n - t^*$ is *detection delay*.

We make the following assumption for M_1 and M_2.

Assumption 7 *Suppose that for some* $0 < K < \infty$*,for any* X*,* $|\log P(X|M_i)| \leq K$ *for* $i = 1, 2$ *and that for some* $0 < V < \infty$ *the variance of the random variable* $V_j = \log P(X_j|X^{j-1} : M_2)/P(X_j|X^{j-1} : M_1)$ *with respect to* $P(X_j|X^{j-1} : M_2)$ *is upper-bounded by* V *for any* j*.*

We give the following theorem on Type 1 and 2 error probabilities for general cases.

Theorem 8 *For DMS as a change-point detector, we have*

$$\text{Type 1 error probability } \leq 2^{-f(n,t^*)}. \tag{9}$$

Let us define the Kullback-Leibler divergence *(the KL-divergence) between* $P(X^h|x^{t^*} : M_2)$ *and* $P(X^h|x^{t^*} : M_1)$ *by*

$$\begin{aligned} &D_h(M_2||M_1)|_{x^{t^*}} \\ &\stackrel{\text{def}}{=} \sum_{X_{t^*+1}^n} P(X_{t^*+1}^n|x^{t^*} : M_2) \log \frac{P(X_{t^*+1}^n|x^{t^*} : M_2)}{P(X_{t^*+1}^n|x^{t^*} : M_1)}. \end{aligned}$$

Under Assumption 7, if $D_h(M_2||M_1)|_{x^{t^*}} > f(n, t^*)$ *holds, for some* $0 < C < \infty$*, we have*

$$\text{Type 2 error probability } \leq 2\exp\left(-Ch\beta_h^2\right), \tag{10}$$

where

$$\beta_h \stackrel{\text{def}}{=} \frac{1}{h}\left(D_h(M_2||M_1)|_{x^{t^*}} - f(n, t^*)\right), \tag{11}$$

Theorem 8 shows that Type 1 error probability for DMS is always upper-bounded by the exponential in the negative $f(n,t^*)$, which is determined by only the code-lengths for model transition. We also see that Type 2 error probability for DMS decays in order $O(\exp(-h\beta_h^2))$, where the exponent factor depends on the code-length for model transition as well as the KL-divergence between M_2 and M_1. The larger the KL-divergence minus $f(n,t^*)$ is, the smaller Type 2 error probability is. The larger $f(n,t^*)$ is, the smaller Type 1 error probability is while the larger Type 2 error probability is. The balance between Type 1 and 2 error probabilities depends on how to estimate model transition probability distributions. We have the following corollaries for the respective model transition estimators.

Corollary 9 *Let the values of $f(n,t_*)$ as in (8) for SM estimator, KT estimator and W estimator be $f^{SM}(n,t^*)$, $f_{KY}(n,t^*)$,and $f^W(n,t^*)$, respectively. Then they are given as follows:*

$$f^{SM}(n,t^*) = \log Z_\infty t^{*(1+\epsilon)} + \log\left\{\left(\frac{h+1}{h+1+\epsilon}\right)\left(\frac{h+1}{n}\right)^\epsilon\right\},$$

$$f^{KT}(n,t^*) = \log(2(t^*+h)-1),$$

$$f^W(n,t^*) = \log\frac{(n-1/2)_h h!}{n_h(h-1/2)_h} + \log(2t^*-1),$$

where $(n-1/2)_h = (n-1/2)(n-3/2)\cdots(t^*+1/2)$ and $n_h = n(n-1)\cdots(t^*+1)$.

We may see that for fixed t^*, for sufficiently large h for sufficiently small $\epsilon > 0$,

$$f^{KT}(n,t^*) > f^{SM}(n,t^*) > f^W(n,t^*). \tag{12}$$

This implies that Type 1 error probability becomes small in this order while Type 2 error probability becomes large in this order.

4. DATA COMPRESSION WITH DMS

4.1. Data Compression

When we apply DMS of Definition 4 into data compression, we have the following theorem on its total code-length:

Theorem 10 *For any x^n, the total code-length for DMS, which we denote as $\ell(x^n)$, is upper-bounded as follows:*

$$\ell(x^n) \le \min_m \min_{t_0,\ldots t_m} \min_{M(0),\ldots M(m)} \Bigg\{\log|\mathcal{M}| + F(n,m) + \sum_{j=0}^{m}\sum_{t=t_j+1}^{t_{j+1}} \big(-\log P(x_t|x^{t-1}:M_t)\big)\Bigg\}, \tag{13}$$

where $t_0 = 0 < t_1 <, \ldots, < t_m < t_{m+1} = n$ denote change-points, m is the number of change-points, $M(j) \in \mathcal{M}$ is the model at $[t_j, t_{j+1})$ $(i = 0,\ldots,m)$, and the minimum is taken under the condition that $|M(j) - M(j+1)| \le 1$ $(j = 0,\ldots,m-1)$. $F(n,m)$ is code-length for a model sequence $M(0)..M(0)M(1).....M(m)$. For SM estimator, KT estimator, and W estimator, we denote $F(n,m)$ as $F_{SM}(n,m)$, $F_{KT}(n,m)$, and $F_W(n,m)$, respectively. They are expanded as follows:

$$F_{SM}(n,m) = m\log\frac{n}{m} + \epsilon(m+1)\log\frac{n}{m+1} + (m+1)\log(1+\epsilon) - m\log\frac{\epsilon}{2},$$

$$F_{KT}(n,m) = (n-1)H\left(\frac{m}{n-1}\right) + \frac{1}{2}\log(n-1) + (m+1)\log 2,$$

$$F_W(n,m) = \frac{3m}{2}\log\frac{n}{m} + \frac{1}{2}\log n + (2m-1)\log 2 + m,$$

where $H(x) = -x\log x - (1-x)\log(1-x)$.

For each m, for any sufficiently large n, for sufficiently small $\epsilon > 0$, the following relation holds among SM, KT, and W:

$$F_{SM}(n,m) < F_{KT}(n,m) < F_W(n,m). \tag{14}$$

4.2. Learning PSMSs

Let $\mathcal{X}$ be either discrete or continuous. Let $\mathcal{F} = \{p(x;\theta) : \theta \in \Theta\}$ be a parametric class of probability distributions (or probability mass functions) where Θ is a parameter space. We suppose that each x_t of $x^n = x_1\ldots x_n \in \mathcal{X}^n$ is independently generated according to a class of probability distributions with $m+1$ piecewise constant parameters as follows:

$$\begin{cases} x_t \sim p(x;\theta(0)) & (1 \le t \le t_1), \\ x_t \sim p(x;\theta(1)) & (t_1+1 \le t \le t_2), \\ \quad\vdots & \\ x_t \sim p(x_t;\theta(m)) & (t_m+1 \le t \le n), \end{cases} \tag{15}$$

where $0 < t_1 < t_2 < \cdots < t_m < n$ $(t_0 = 0,\ t_{m+1} = n)$ is a sequence of change-points and each $\theta(j) \in \Theta$ $(j = 0,\ldots,m)$ and $\theta(j) \ne \theta(j+1)$ $(j = 0,\ldots,m-1)$. We call such a source a *piecewise stationary memoryless source* (PSMS) [7],[9].

We consider any lossless data compression algorithm $\mathcal{A}$, which takes as input x^n and outputs a lossless compressed data sequence. We denote the total code-length for x^n using $\mathcal{A}$ as $\mathcal{L}_\mathcal{A}(x^n)$. We define as a measure for the goodness of $\mathcal{A}$ the *expected redundancy* as follows:

Definition 11 For any lossless data compression algorithm $\mathcal{A}$, for a given PSMS as in (15), we define the *expected redundancy* for $\mathcal{A}$ by

$$\mathcal{R}^n_\mathcal{A} \stackrel{\text{def}}{=} \mathbb{E}\left[\mathcal{L}(x^n) - \sum_{j=0}^{m}\sum_{t=t_j+1}^{t_{j+1}} \big(-\log p(x_t;\theta(j))\big)\right],$$

where the expectation is taken with respect to (15).

Merhav[7] derived the following lower bound on the expected redundancy.

Theorem 12 [7] *Suppose that the domain $\mathcal{X}$ is finite. Supposing that each datum is independently generated according to almost any PSMS with fixed m as the number of change-points and fixed k as the degrees of freedom of each parameter, and under other some conditions for any $\varepsilon > 0$ and sufficiently large n, we have*

$$\inf_\mathcal{A} \mathcal{R}^n_\mathcal{A} \ge (1-\varepsilon)\left(\frac{k(m+1)}{2}\log n + m\log n\right). \tag{16}$$

In the case where Θ is 1-dimensional and compact, Kanazawa and Yamanishi[4] applied DMS to develop an algorithm that asymptotically matched (16).Below we introduce their approach. The key ideas of their algorithm are summarized as follows:

1) Discretization of parameter space: For a given positive integer K, we discretize Θ to obtain a finite set of size K. Let us define Fisher information associated with $\mathcal{F}$ and L_I by

$$I(\theta) \stackrel{\text{def}}{=} \mathbb{E}_\theta\left[-\frac{\partial^2\log p(x;\theta)}{\partial\theta^2}\right],\quad L_I \stackrel{\text{def}}{=} \int_{\theta\in\Theta}\sqrt{I(\theta)}\mathrm{d}\theta,$$

respectively. Letting $\delta_I = L_I/(K-1)$ be a discretization scale and $\bar\theta_1 = \theta_{\min}$, we define $\bar\theta_i$ so that

$$\int_{\bar\theta_1}^{\bar\theta_i}\sqrt{I(\theta)}\mathrm{d}\theta = (i-1)\delta_I \quad (i = 2,\ldots,K). \tag{17}$$

We have $\bar{\Theta} = \{\bar{\theta}_1, \ldots, \bar{\theta}_K\}$. We assume that for each interval $\bar{\theta}_i \le \theta \le \bar{\theta}_{i+1}$, either $\mathrm{d}\sqrt{I(\theta)}/\mathrm{d}\theta \le 0$ or $\mathrm{d}\sqrt{I(\theta)}/\mathrm{d}\theta \ge 0$.

2)Settings of model transition probabilities: When the model set is a set of discretized parameters, it may be difficult to assume that the parameter transits to neighbouring ones only as in Definition 1). In that case, we assume according to [4] that the parameter value transits according to the following probabilities:

$$\Pr(i_t \mid i_{t-1}) = \begin{cases} \frac{\alpha}{K-1} & (i_t \neq i_{t-1}), \\ 1-\alpha & (i_t = i_{t-1}). \end{cases} \quad (18)$$

where we set K and α as

$$K = \lfloor \sqrt{n} \rfloor, \quad \alpha = 1/n.$$

Under the above setting Kanazawa and Yamanishi [4] proposed an algorithm for learning PSMSs that takes x^n as input and outputs the parameter sequence $(\bar{\theta}_{i_1}, \ldots, \bar{\theta}_{i_n})$ where $i_1, \ldots, i_n$ are those which attain the DMS criterion. Its performance is summarized in the following theorem:

Theorem 13 [4] *Suppose that each datum is independently drawn according to a PSMS. There exists an algorithm $\mathcal{A}$ for which time complexity is $O(n^{3/2})$ and the expected redundancy satisfies:*

$$\mathcal{R}^n_{\mathcal{A}} < \frac{m+1}{2}\log n + m\log n + \frac{L_I^2}{2} + \log \mathrm{e} + O(n^{-1/2}). \quad (19)$$

The bound (19) implies that the expected redundancy for the algorithm asymptotically matches the lower bound (16).

5. DISTINGUISHABILITY

Let us employ $\mathcal{F} = \{p(x;\theta) : \theta \in \Theta\}$ as a model class of probability distributions (or probability mass functions) where Θ is a 1-dimensional real-valued parameter space. We consider how to discretize Θ to get a finite subset $\overline{\Theta}$. From the argument in Section 4.2(see 17), we see that if we let the discretization scale $\delta = \max_i |\bar{\theta}_i - \bar{\theta}_{i+1}|$ be

$$\delta = O\left(\sqrt{1/n}\right) \quad (20)$$

then we have an upper bound on the expected redundancy which attains Merhav's lower bound. In this sense the discretization scale as in (20) is optimal in the scenario of data compression. It coincides with results in [8],[1].

Meanwhile, let us consider the case where DMS is applied into change-point detection over a discretized parameter set $\overline{\Theta}$. When either SM, KT, W estimator or the uniform model transition probability as in (18) is employed for model transition estimation, we see from Theorem 8 that Type 2 error probability for DMS decreases exponentially with respect to n if

$$\min_{\bar{\theta}(\neq)\bar{\theta}' \in \overline{\Theta}} D(\theta||\theta') > f(n,t^*)/n = O(\log n/n). \quad (21)$$

Note that for any $\theta, \theta' \in \overline{\Theta}$, we have $D(\theta||\theta') = (1/2)I(\theta)\delta^2$, where δ is the discretization scale. If

$$\delta = O\left(\sqrt{\log n/n}\right) \quad (22)$$

then (21) holds. The discretization scale (22) makes the total code-length $(1/2)\log n$ larger than the bound (19). This implies that (22) doesn't lead to optimal data compression. Hence there is a gap between the optimal discretization in the sense of change-detection and that of data compression. Change-detection requires more discriminability over the parameter space than data compression.

6. CONCLUSION

We have applied DMS into the scenarios of change-detection and data compression for time-varying sources. We have analyzed the performance of DMS in the both scenarios and have shown how it is related to model transition estimation. We have argued how to discretize the real-valued parameter space to obtain optimal performance in the both scenarios. It has turned out that change-detection may require more discriminability over the parameter space than data compression.

7. ACKNOWLEDGMENTS

This work was partially supported by MEXT KAKENHI 23240019, Aihara Project, the FIRST program from JSPS, initiated by CSTP, NTT Corporation.

8. REFERENCES

[1] V.Blasubramanian. Statistical inference, Occam's razor and statistical mechanics on the space of probability distributions. *Neural Computation*, 9, No.2, pp:349–368, 1997.

[2] T. van Erven and P.D. Grünwald and S. de Rooij. Catching up faster in Bayesian model selection and model averaging. *Advances in NIPS* 20, 2007.

[3] S. Hirai and K.Yamanishi. Detecting changes of clustering structures using normalized maximum likelihood coding. *Proc. of the eighteenth ACM SIGKDD Int'l.Conf.on Knowledge Discovery in Data Mining(KDD2012)*, 2012.

[4] H.Kanazawa and K.Yamanishi. An MDL-based change-detection with its applications to learning piecewise stationary memoryless sources. *Proc. of IEEE Information Theory Workshop(ITW2012)*, 2012.

[5] J. Kleinberg. Bursty and hierarchical structure in streams. *D. M. K. D.*, vol. 7, pp. 373–397, Nov. 2003.

[6] R. E. Krichevsky and V. K. Trofimov. The performance of universal encoding. *IEEE Trans. Inf. Theory*, 27:199–207, 1981.

[7] N. Merhav. On the minimum description length principle for sources with piecewise constant parameters. *IEEE Trans. Inf. Theory*, vol. 39, pp. 1962–1967, Nov. 1993.

[8] J. Rissanen. *Information and Complexity in Statistical Modeling*, Springer, 2007.

[9] G. I. Shamir and D. J. Costello, Jr. Asymptotically optimal low-complexity sequential lossless coding for piecewise-stationary memoryless sources—Part I: The regular case. *IEEE Trans. Inf. Theory*, vol. 46, pp. 2444–2467, 2000.

[10] G. I. Shamir and N. Merhav. Low complexity sequential lossless coding for piecewise stationary memoryless sources. *IEEE Trans. Inf. Theory*, Vol.45,pp:1498–1519,1999.

[11] F. M. J. Willems.Coding for a binary independent piecewise identically-distributed source. *IEEE Trans. Inf. Theory*,Vol.42, pp:2210–2217, 1996.

[12] F. M. J. Willems and F. Casadei. Weighted coding methods for binary piecewise memoryless sources. *Proc. of 1995 IEEE ISIT*, p.323,1995.

[13] K. Yamanishi and Y. Maruyama. Dynamic syslog mining for network failure monitoring. *Proc. of KDD2005*, pp: 499-508, ACM Press, 2005.

[14] K. Yamanishi and Y. Maruyama. Dynamic model selection with its applications to novelty detection. *IEEE Trans. Inf. Theory*,IT 53(6)：2180-2189, 2007.