On Detection of Changes in Attribute Data

Emmanuel Yashchin

IBM Research

T. J. Watson Research Center
Yorktown Heights, NY 10598

2003 Quality and Productivity Research Conference, Yorktown Heights, May 2003

Overview

- Categorical Data Models
- Monitoring Schemes
- Examples
- Generalizations

Categorical Data Models

Example: Monitoring level of contamination (IC manufacturing)

Areas of interest: A1, A2, A3

Particle categories: ML, OC, IC, OR

Counts by Particle Category (Collected Daily)

	Metal	Organic	Inorganic	Other	
Area 1	24	12	14	10	60
Area 2	20	13	7	9	49
Area 3	12	14	9	6	41
	56	39	30	25	150

The Model

Observations: Sequence of contingency tables $\{B_t\}$, $t = 1, 2, \ldots$.

Table type: two-way (I rows, J columns)

n_t = number of particles in a table B_t ($E(n_t) = \lambda_t$).

$n_{ij}, n_{i.}, n_{.j}$ = number of particles in cell (i, j), row i, column j.
$E(n_{ij}, n_{i.}, n_{.j}) = \lambda_{ij}, \lambda_{i.}, \lambda_{.j}$.

$p_{ij}, p_{i.}, p_{.j}$ = nominal prob. that a particle is in cell (i, j), row i, col. j

$\pi_{ij}, \pi_{i.}, \pi_{.j}$ = actual prob. that a particle is in cell (i, j), row i, col. j

The Monitored Parameters

Denote: $\tilde{p}_{ij} = p_{ij}/p_{i.}$ and $\tilde{\pi}_{ij} = \pi_{ij}/\pi_{i.}$.

Conditional Row Distributions: $\tilde{p}_i = (\tilde{p}_{ij}, j = 1, \ldots, J)$ (nominal) and $\tilde{\pi}_i = (\tilde{\pi}_{ij}, j = 1, \ldots, J)$ (actual).

Definition: $FCR\{\tilde{p}_{ij}\}$ = family of two-way distributions with conditional distributions in rows given by $(\tilde{p}_i, i = 1, \ldots, I)$.

Area Parameters: $\tilde{p}_{ij}, \lambda_{i.}$ and d_i (Kullback - Leibler distance),

$$d_i = I(\tilde{p}_i, \tilde{\pi}_i) = \sum_j \tilde{p}_{ij} ln(\tilde{p}_{ij}/\tilde{\pi}_{ij}).$$

Global Parameters

Distance d of $\{\tilde{\pi}_{ij}\}$ from $FCR\{\tilde{p}_{ij}\}$ family:

$$d = -ln\left[\sum_i \pi_{i.} exp(-d_i)\right].$$

Coefficent of Association, η

$$\eta = 1 - \frac{\sum_{i,j} \pi_{ij} ln(\pi_{ij}/\pi_{i.})}{\sum_j \pi_{.j} ln(\pi_{.j})}$$

Overall Particle Rate:

$$\lambda = \sum_i \lambda_{i.}.$$

Monitoring: Basic Approach

- Decompose data stream into sequences of sufficient statistics
- Assign priorities to monitored parameters
- For each parameter establish acceptable / unacceptable levels
- Specify acceptable rate of false alarms
- Design Regenerative Likelihood Ratio (RLR) control schemes
- Test and deploy

Likelihood Ratio (LR) Approach

Ω_0 – acceptable region Ω_1 – unacceptable region

Current observation: X_T

Log-likelihood: $L_m(\theta) = \ln f_\theta(X_{T-m+1}, \ldots, X_{T-1}, X_T)$

Max log-likelihoods: $L^*_{m0} = \max_{\theta \in \Omega_0} L_m(\theta)$, $L^*_{m1} = \max_{\theta \in \Omega_1} L_m(\theta)$

Score: $D^*_m = L^*_{m1} - L^*_{m0}$

LR approach: Select signal level $h > 0$. Signal at time T if $D^*_m > h$ for some $m \geq 1$.

Enhancements: RLR scheme, providing a mechanism for discarding old history

Regenerative LR approach (RLR)

Regeneration point: All information prior to this point is discarded

At time T: Last regeneration point recorded M_T observations ago

RLR approach:

1. Signal at time T if $D_m^* > h$ for some $1 \le m \le M_T$.

2.
   - If $D_m^* \le 0$ *for every* $1 \le m \le M_T$, declare T a new regeneration point
   - Otherwise, denote by m_T the maximal value of m in $[1, M_T]$ for which $D_m^* > 0$, and declare $T - m_T$ a new regeneration point.
     Note: if $D_m^* > 0$ for $m = M_T$, keep the current regeneration point

RLR Control Chart:

1. Define the value of scheme s_T at time T by

$$s_T = \max_{1 \le m \le M_T} D_m^*$$

2. Plot s_T on a control chart

3. Trigger a signal at time T if $s_T > h$.

4. If $s_T \le 0$ declare T the new regeneration point. Otherwise, find m_T and declare $T - m_T$ a new regeneration point.

The Control Schemes: Basic Parameters

Monitoring λ:

Acceptable Region: $\lambda \leq \underline{\lambda}$ **Unacceptable Region:** $\lambda \geq \overline{\lambda}$ ($\underline{\lambda} < \overline{\lambda}$).

Upper LR scheme for λ: Trigger a signal at time T if for some $m \geq 1$

$$k_\lambda \leq \hat{\lambda}_{(m)} \leq \overline{\lambda} \quad \text{and} \quad m(\ln \overline{\lambda} - \ln \underline{\lambda})(\hat{\lambda}_{(m)} - k_\lambda) > h_\lambda \quad \text{or}$$
$$\hat{\lambda}_{(m)} \geq \overline{\lambda} \quad \text{and} \quad m[\hat{\lambda}_{(m)}(ln\hat{\lambda}_{(m)} - ln\underline{\lambda}) - (\hat{\lambda}_{(m)} - \underline{\lambda})] > h_\lambda,$$

where $\hat{\lambda}_{(m)} = (1/m)\sum_{T-m+1}^{T} n_t$, h_λ is a *signal level* and k_λ is the *reference value*, $k_\lambda = (\overline{\lambda} - \underline{\lambda})/(ln\overline{\lambda} - ln\underline{\lambda})$.

Monitoring Functions of Parameters

Monitoring d:

Acceptable Region: $d \leq \underline{d}$ **Unacceptable Region:** $d \geq \overline{d}$ ($\underline{d} < \overline{d}$).

Computing L^*_{m0} and L^*_{m1}:

(a) If $\hat{d}_{(m)} \leq \underline{d}$ set $L^*_{m0} = L^*_m$ and $L^*_{m1} = \max_{d(\theta)=\overline{d}} L_m(\theta)$.

(b) If $\hat{d}_{(m)} \geq \overline{d}$ set $L^*_{m1} = L^*_m$ and $L^*_{m0} = \max_{d(\theta)=\underline{d}} L_m(\theta)$.

(b) Else set $L^*_{m0} = \max_{d(\theta)=\underline{d}} L_m(\theta)$ and $L^*_{m1} = \max_{d(\theta)=\overline{d}} L_m(\theta)$.

Note: in case (a) $D^*_m < 0$, so no effort is needed.

Monitoring η: Use a similar technique.

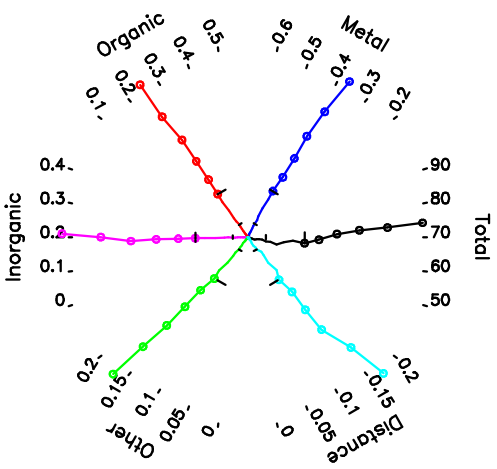
A simplified scheme: Geometric Cusum (GC):

Upper GC scheme for $\mathbf{d}$: Trigger an out of control signal when the process s_t defined by

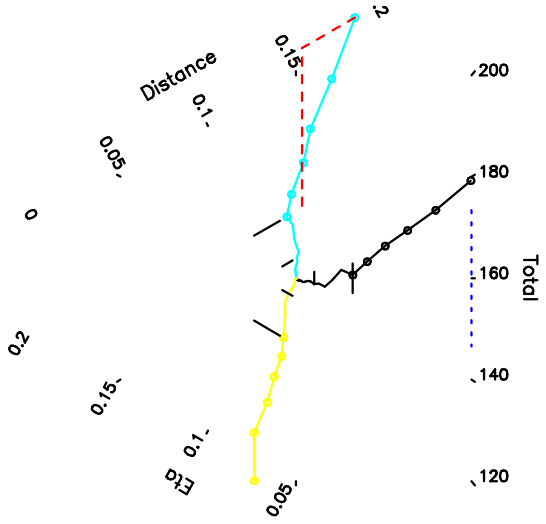
$$s_0 = 0, \quad s_t = max\{0, \gamma s_{t-1} + n_t(\sqrt{\hat{d}_t} - k_d)\}, \quad t = 1, 2, \ldots,$$

(with $\gamma \approx 0.8$ and $k_d \approx 0.5(\sqrt{\underline{d}} + \sqrt{\overline{d}})$) exceeds threshold h_d.

Example: based on data for last 25 days

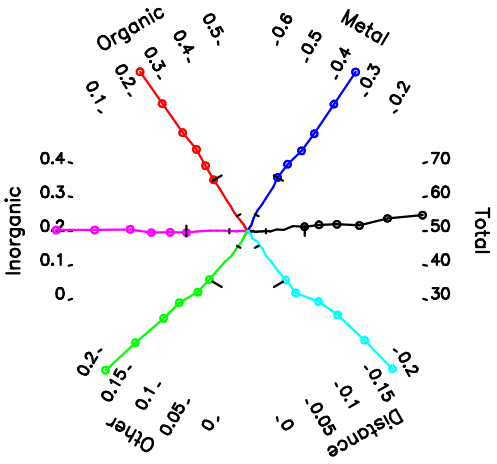


Area 1 (5 day counts = 400/350)

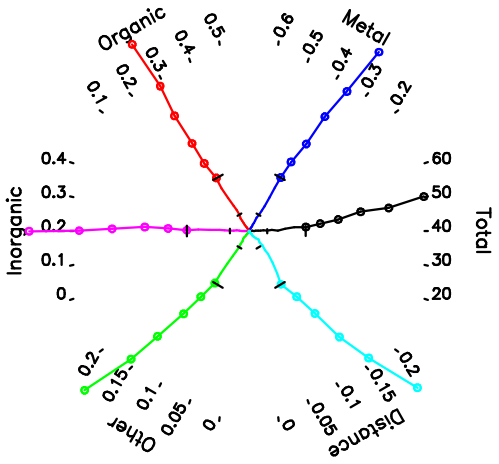


Global Parameters (5 day counts = 940/800)

April 25, 2003



Area 2 (5 day counts = 274/250)



Area 3 (5 day counts = 266/200)

Conclusions

- Monitoring methods based on use of Regenerative Likelihood Ratios in conjunction with change-point models are promising for attribute data.

- Generalizations are available for multi-way classifications.

- Open question: RLR optimality properties?