Distributed Clustering and Change Detection

Michèle Sebag

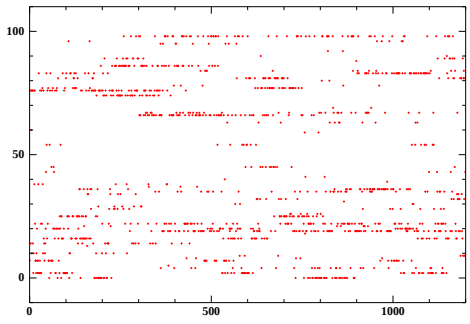
KDUBiq Summer School, Porto-2008
http://www.lri.fr/∼sebag/KDU

1 Distributed Clustering

Consider a set of computing nodes. Each node:

- receives data items from a datastream, DataStream.N
- clusters them,
- periodically transmits its best cluster to another node,
- periodically receives some cluster description from another node. Transmitted.N

Figure 1: Non-stationary distribution: Index of the cluster vs time



1.1 Assumptions

The datastream involves N clusters, sampled according a non-stationary distribution. Each node can at most construct $K < N$ clusters.

1.2 Steps

1. Achieve K-means with change point detection (Section 2).
2. Exploit the clusters provided by other nodes. Beware, they can be irrelevant.
3. Compare the total distortion with the oracle one.

2 Change Point Detection: Page Hinkley

PH: parameters λ, controls the false alarm rate
δ, tolerance

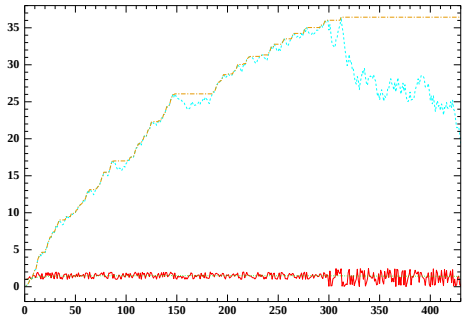
series $p_1, \ldots, p_t$

$$\bar{p}_t = \frac{1}{t} \sum_{\ell=1}^{t} p_\ell$$

$$m_t = \sum_{\ell=1}^{t} p_\ell - \bar{p}_\ell + \delta$$

$$M_t = max\{m_\ell, \ell = 1...t\}$$

$$PH_t = (M_t - m_t > \lambda)$$



The Page-Hinkley test: $p_t, \bar{p}_t, m_t, M_t$ vs time

References

Mouss, H., Mouss, D., Mouss, N., & Sefouhi, L. (2004). Test of page-hinkley, an approach for fault detection in an agro-alimentary production system. *5th Asian Control Conference* (pp. 815– 818).

Moustakides, G. (1986). Optimal stopping times for detecting changes in distributions. *Anals of Statistics*, *14*, 1379–1387.

Page, E. (1954). Continuous inspection schemes. *Biometrika*, *41*, 100–115.