High-Dimensional Incremental Divisive Clustering under Population Drift

Nicos Pavlidis

Inference for Change-Point and Related Processes

joint work with David Hofmeyr and Idris Eckley

Clustering: A central problem in pattern recognition

The process of partitioning a set of data objects into disjoint groups (clusters), so that objects of the same cluster are more similar to each other, with respect to a given similarity measure, compared to objects in different clusters

- Improve understanding of data: e.g. Document clustering
- First step for different purposes: e.g. Market Segmentation

- Not unique definition of **true cluster**
- Progress in automated data acquisition and storage technologies generates novel challenges to clustering

- "distances between points become relatively uniform" (Beyer et al 1999)

$$\lim_{d \to \infty} \frac{\text{MaxDist} - \text{MinDist}}{\text{MinDist}} = 0$$

- *things are sometimes not as bad as they might seem* (Steibach et al 2003)
- **Local** rather than global feature relevance / correlation

High-Dimensional Clustering

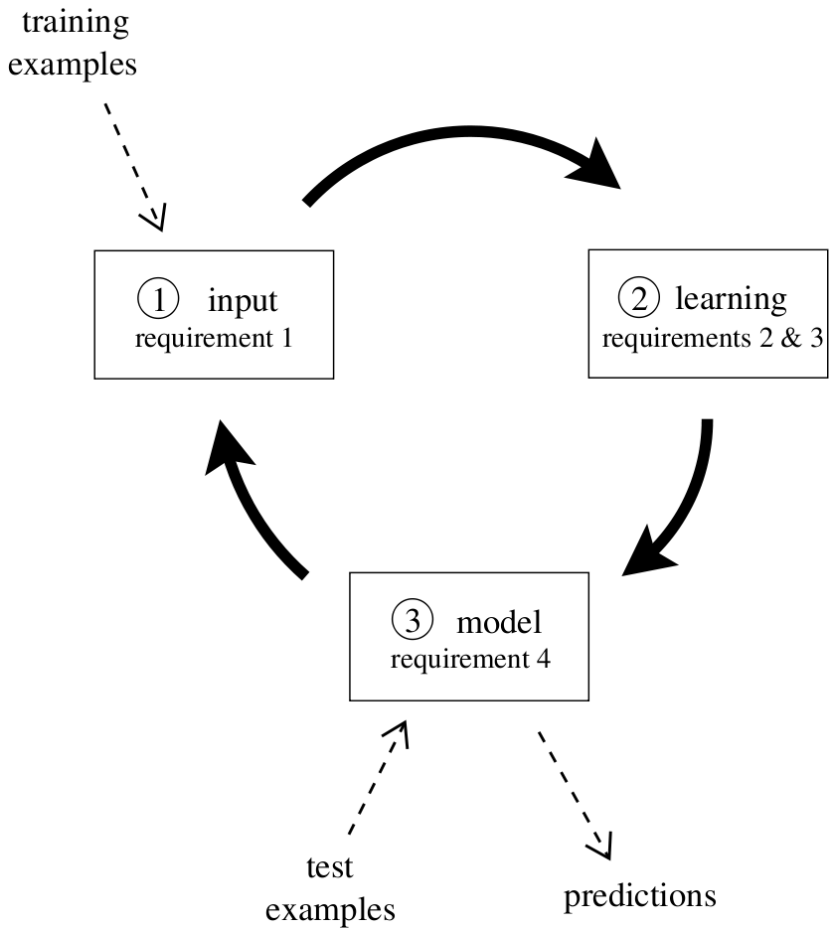
- Search for the relevant subspaces
- Detection of final clusters

- Very active research area over the last 10-15 years
- With very few exceptions all approaches assume unlimited access to the data and static environment

Data Stream

Process generating data sequentially at high frequency relative to available processing or storage capabilities.

1. Process one example at a time
   - Once inspected or ignored it is discarded
2. Limited memory usage
3. Limited time per example
4. Anytime: Supply a model that can be used for prediction at any time



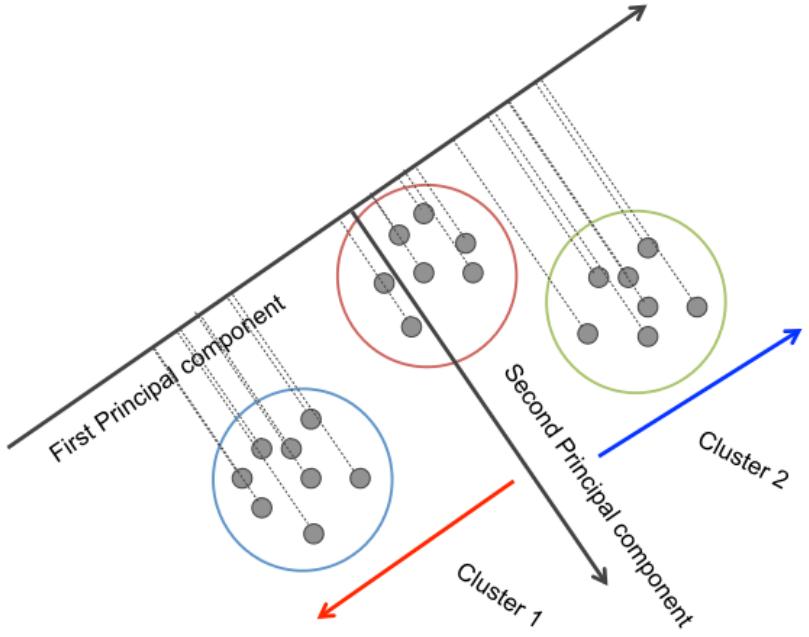
Methods for High-Dimensional Clustering Data-Streams

- Sliding windows: Very sensitive to window size, unstable and inconsistent results over time
- CluStream (Aggarwal et al., 2003):
  - *Microcluster*: Accumulates first and second order spatial and temporal information
  - Executes static offline clustering (k-means) on micro-clusters rather than original data
  - Handles non-stationarity using windows
- HPStream (Aggarwal et al., 2004):
  - High-dimensional "version" of CluStream
  - Each cluster defined in an axis parallel subspace (ignores dimensions in which clusters are less cohesive)
  - Handles non-stationarity through exponential decay – forgetting
- Both require offline initialisation and specification of the number of clusters

Background: Projected Divisive Clustering

Principal Direction Divisive Partitioning (PDDP) (Boley, 1998):

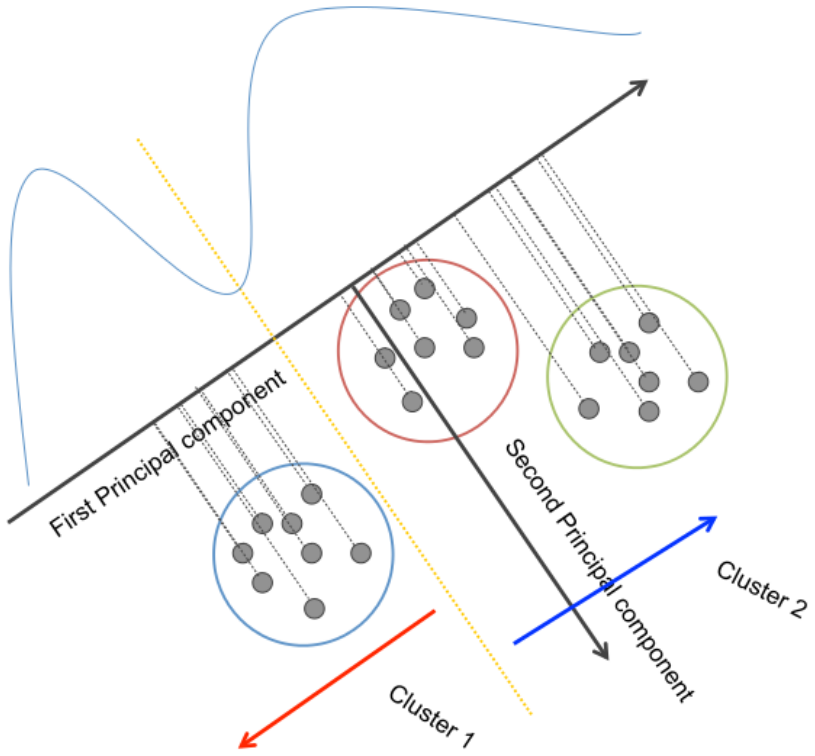
- Project onto first principal component
- Split point at mean projection
  - Well separated clusters can be well split at combined mean
  - A cluster that is "split" can be separated in subsequent iterations
- Termination criterion requires domain specific knowledge



Background: Projected Divisive Clustering

dePDDP (Tasoulis et al., 2010):

- Split point at lowest local minimum in projected density estimate (KDE)
  - Hyperplane with lowest density integral: Avoid splitting clusters
- Modality-based cluster definition: terminate when all marginal distributions are unimodal



Projected Divisive Clustering

Advantages

- Challenging problem of clustering HD data reduced to collection of simple subproblems
- Exploits sparsity: Restriction to convex polyhedra doesn't compromise accuracy
- Incremental PCA algorithms with guaranteed convergence and $\mathcal{O}(1)$ complexity
- Incremental (and adaptive) kernel density estimation methods with bounded memory

Difficulties

- When should a cluster be separated?
- Continually updating hierarchical models compounds instability
- Dealing with non-stationary data – Population drift

Separability of a Streaming Sample

- The Dip Test (Hartigan & Hartigan, 1985) tests for multimodality of a 1D sample
  - Finds closest distribution function with unimodal density to the empirical cdf of a sample (supremum norm)
  - Requires a full sample
- Approximate sample by m compact intervals and associated counts
- **Lemma**: Dip Statistic computed on such sample lower bounds the true $\Rightarrow$ Avoids introducing erroneous splits
  - Streaming Dip Statistic computed in $\mathcal{O}(m)$ time
  - Don't have optimal way of selecting intervals, but observed approximation is better when distribution is multimodal
- After split projection direction and split point remain fixed

Accuracy of Projected Sample

- Shifting projection $\Rightarrow$ non-stationary sample
- Forgetting factor $\lambda \in [0, 1]$

$$w_{t,i} = (1-\lambda)^{t-i}, \quad W_t = \sum_{i=1}^{t} w_{t,i}$$

$$\hat{f}_t(p) = \frac{1}{W_t} \sum_{i=1}^{t} \frac{w_{t,i}}{h_i} K\left(\frac{p - p_i}{h_i}\right)$$

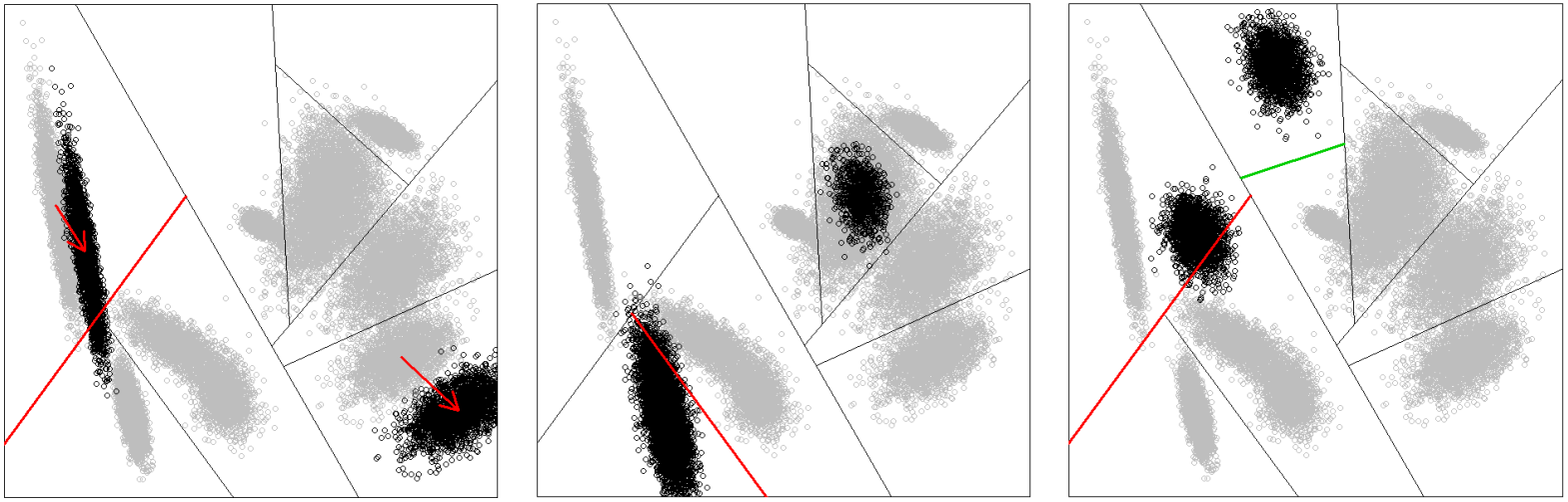
- Updates of projection direction used to adapt λ

$$\lambda_{t+1} = \gamma\lambda_t + (1-\gamma) \arccos\left(e_{t+1}^T e_t\right)$$

- **Lemma**: $e_t \to e \Rightarrow \lambda_t \to 0$

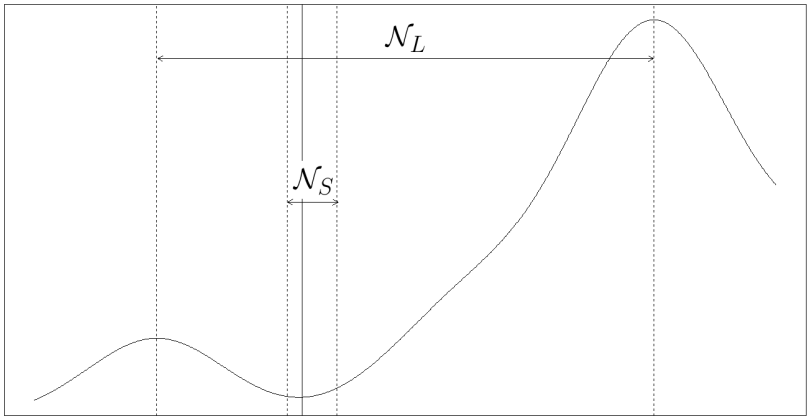
Handling Non-Stationarity

- Numerous types of change possible: Cluster creation/ deletion, Abrupt or gradual shift, Change of shape
- **Intuition**: Hierarchy "correctly" partitions clusters



- Changes rendering part of the hierarchy inaccurate are those which invalidate splitting rule at internal nodes
  - Not all change renders the model obsolete
  - Splitting leaf nodes does not invalidate hierarchy

- Determine whether split point is local minimum



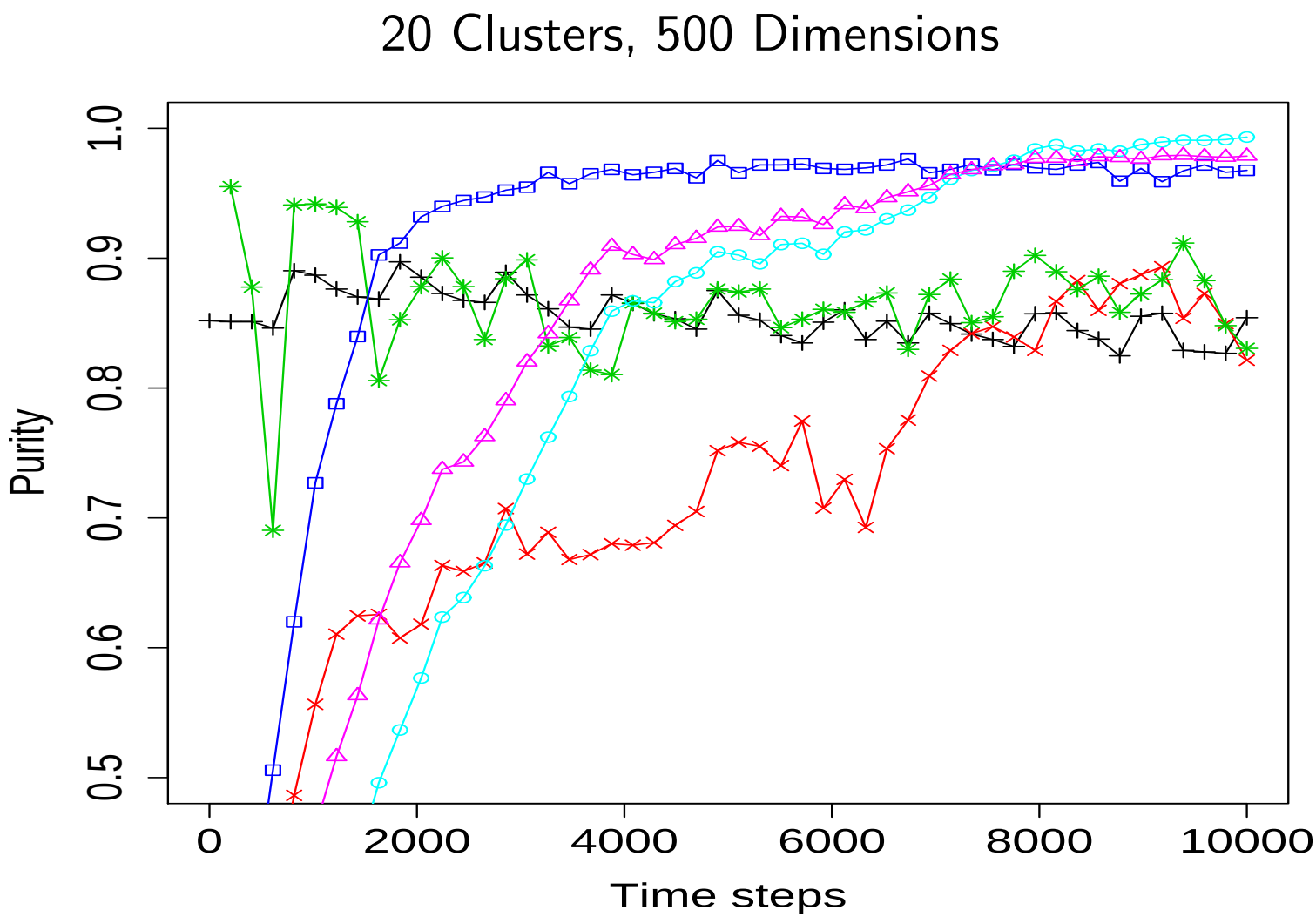
- Bernoulli CUSUM:
  - $\mathcal{N}_S$ contains proportion β of $\mathcal{N}_L$ on either side
  - p_0 estimated using KDE, $p_1 = \beta$

$$B = \begin{cases} 1, & \text{if } x_t \in \mathcal{N}_S \\ 0, & \text{otherwise} \end{cases}$$

$$S_t = \max\{0, S_{t-1}\} + B + \log\left(\frac{1-p_1}{1-p_0}\right) \log\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)^{-1}$$

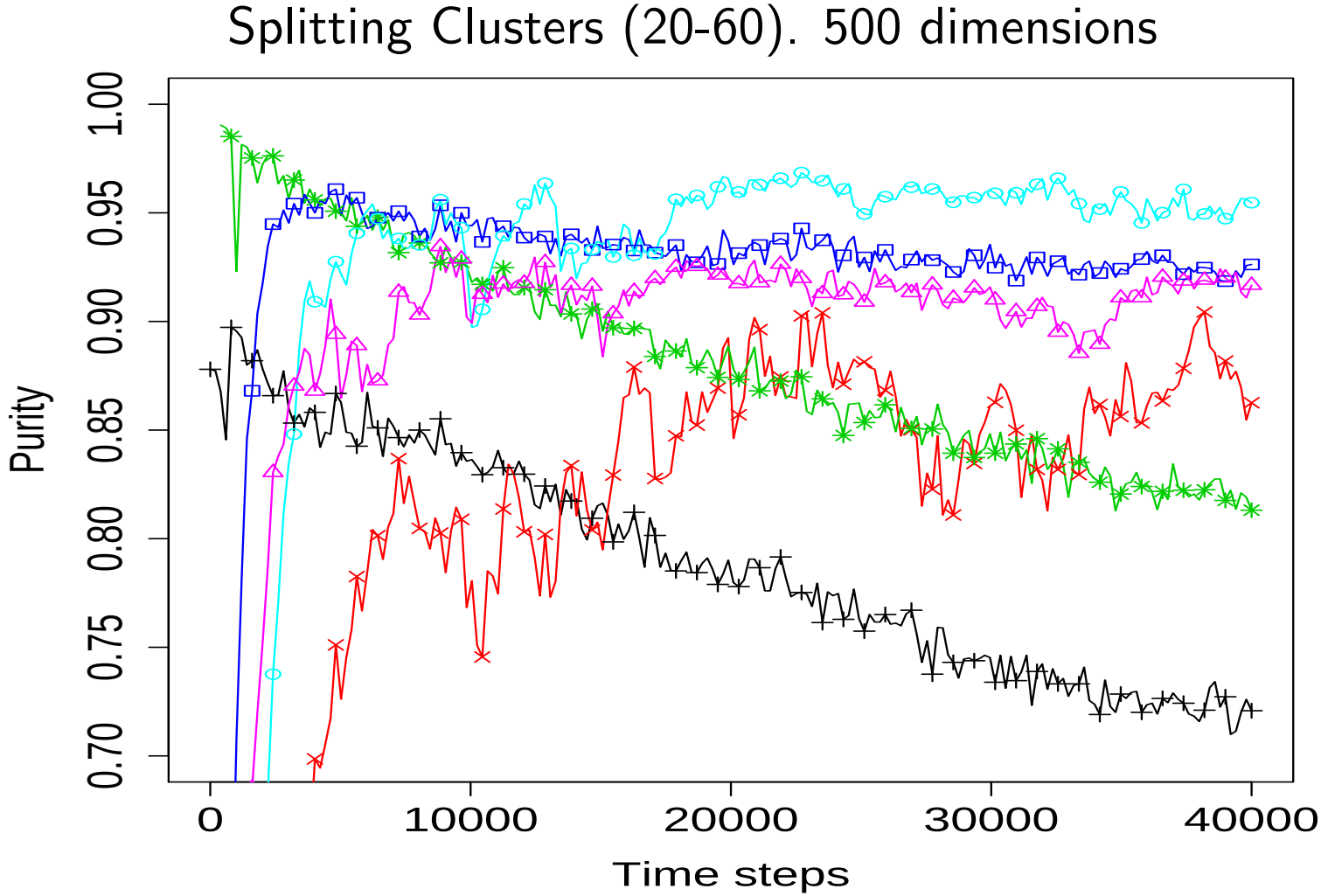
- Corrects erroneous splits due to sequential observation of data

Simulated Datasets: Static



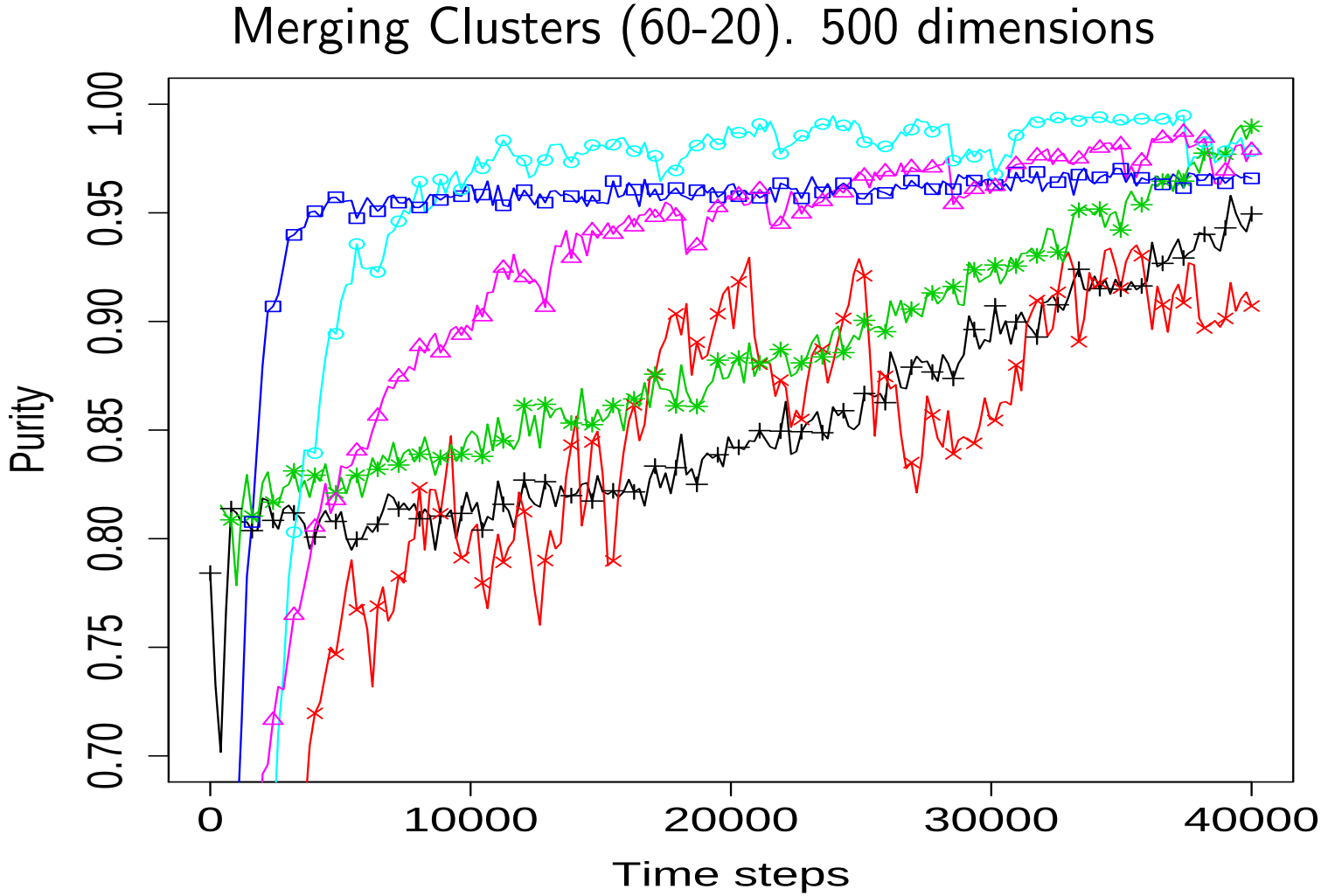
HPStream (+), SPDC (×), CluStream (∗), FP(0) (□), FP (o), FP(B) (△)

Non-Stationary Results



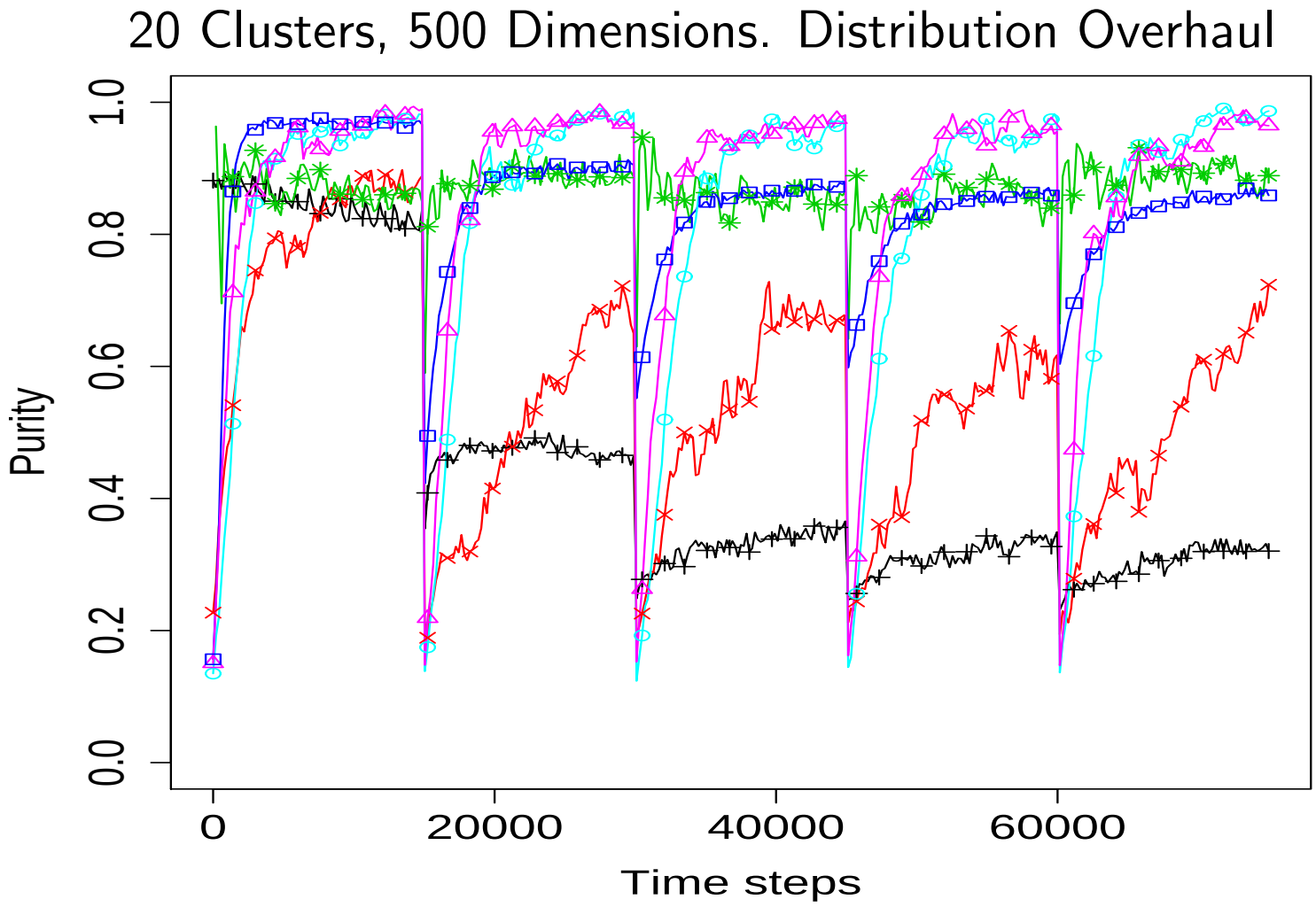
HPStream (+), SPDC (×), CluStream (∗), FP(0) (□), FP (o), FP(B) (△)

Non-Stationary Results



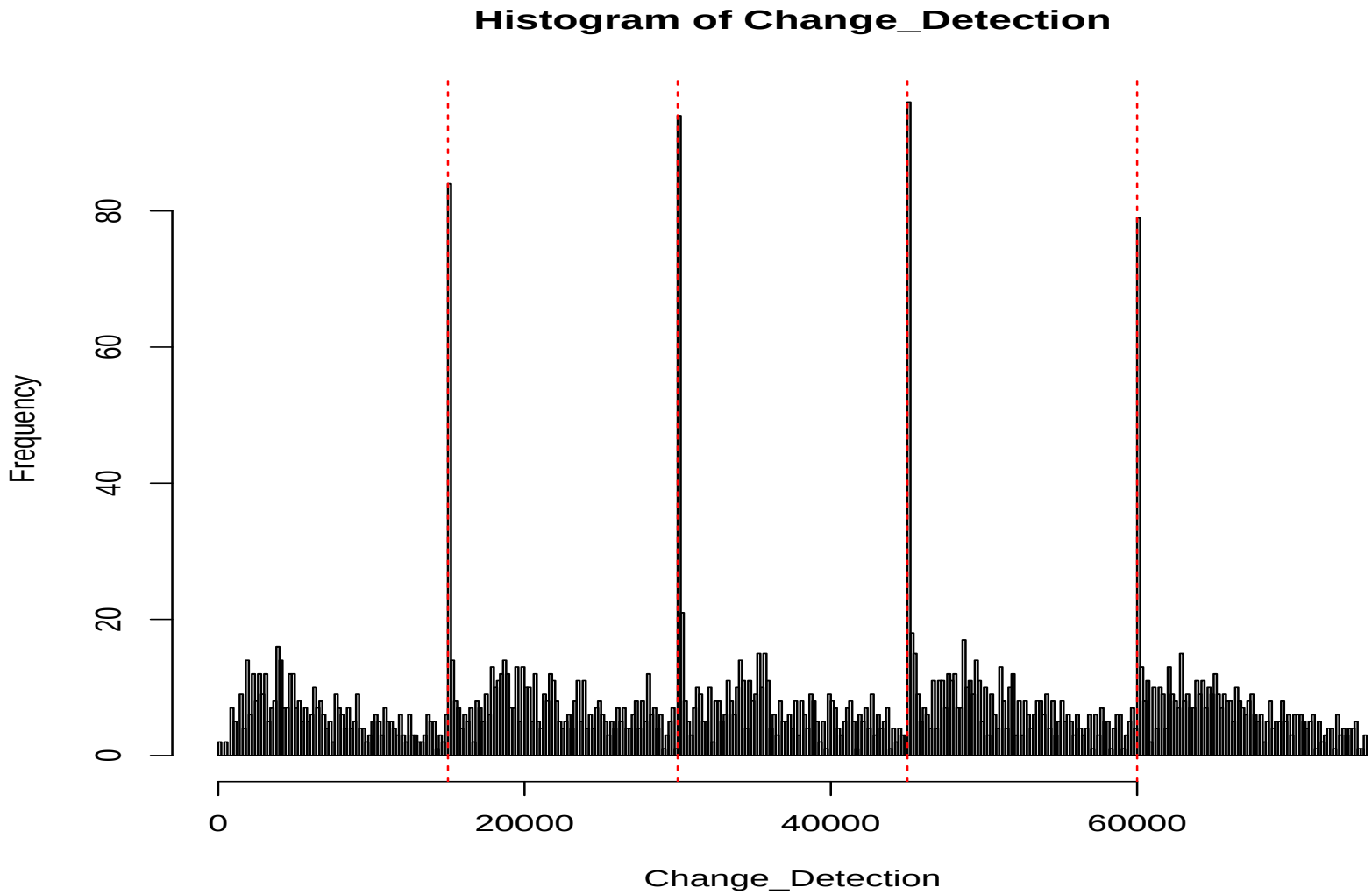
HPStream (+), SPDC (×), CluStream (*), FP(0) (□), FP (o), FP(B) (△)

Non-Stationary Results



HPStream (+), SPDC (×), CluStream (∗), FP(0) (□), FP (o), FP(B) (△)

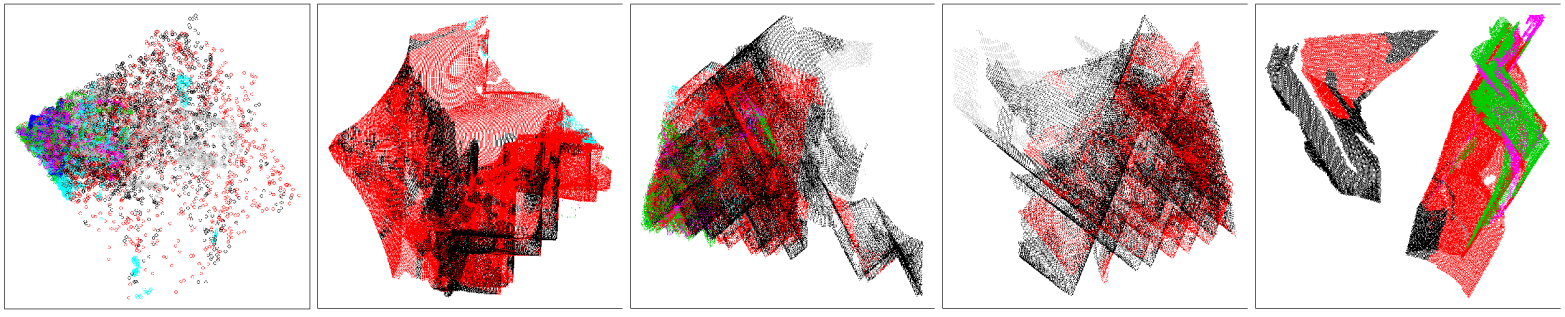
Non-Stationary Results

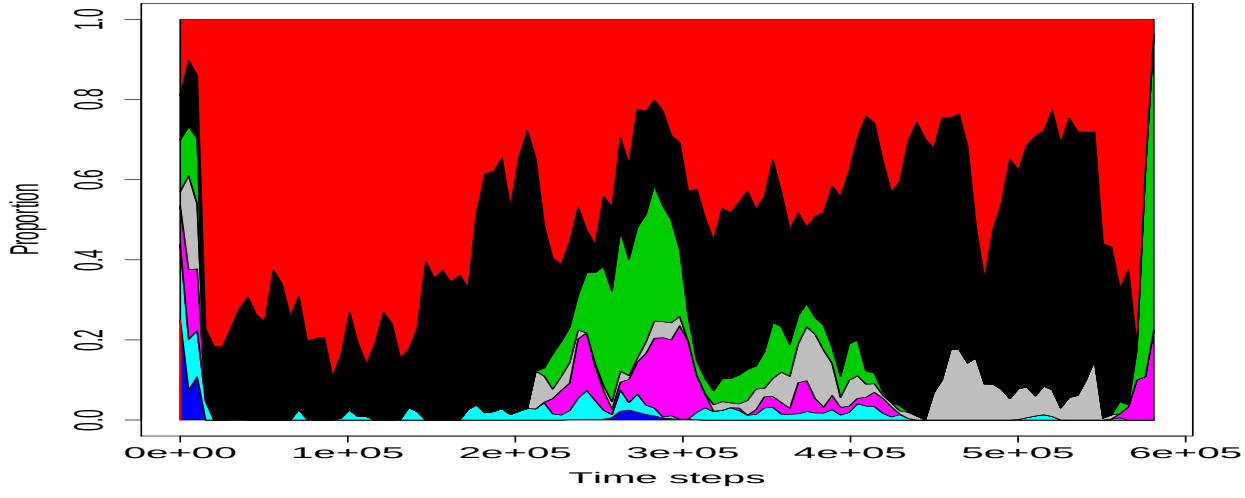


Forest Cover Type Data

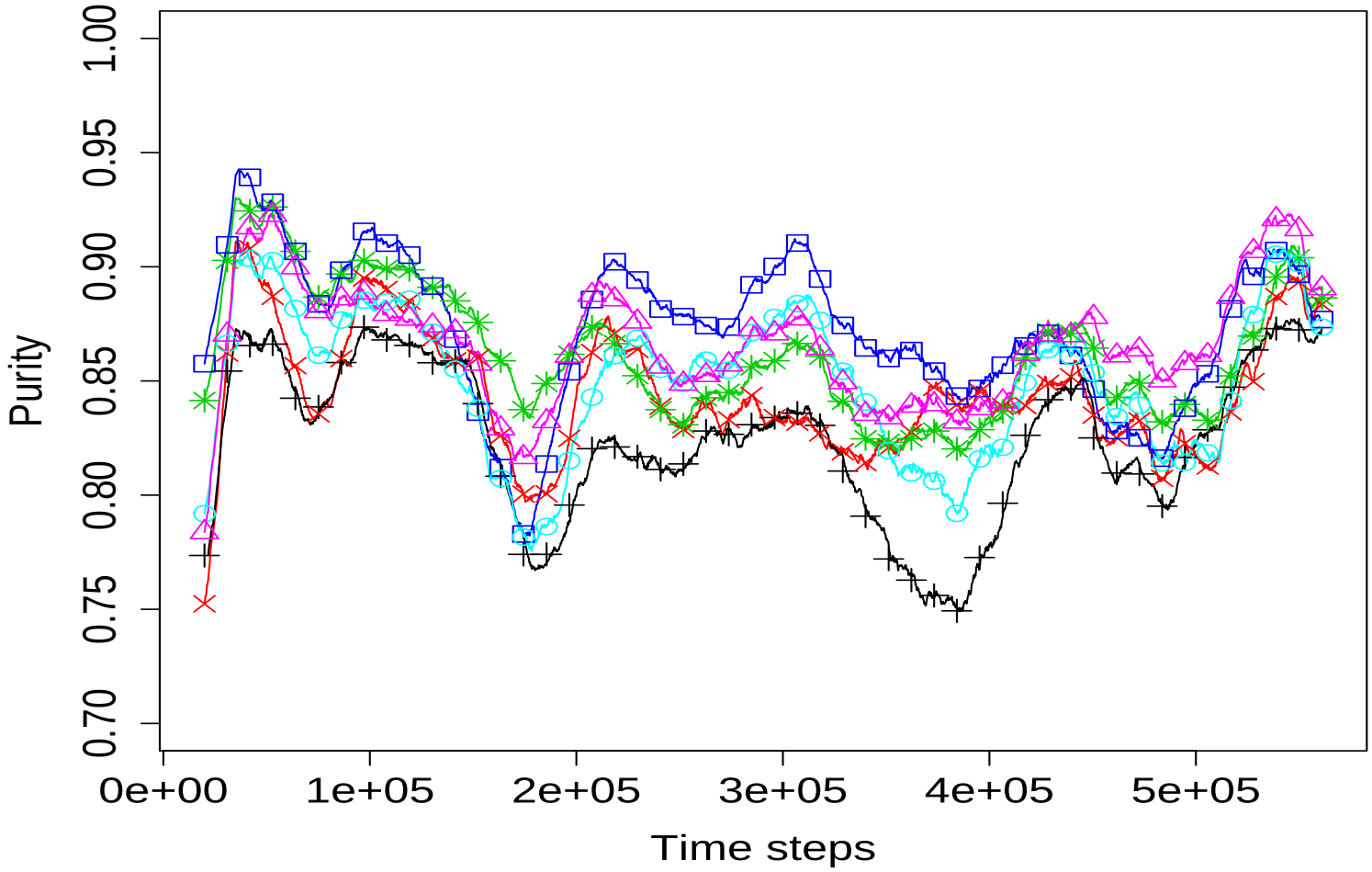
- UCI: 581,012 obs, 7 clusters, 10 num features

Figure: Forest Cover Type Divided Into 5 Segments





Forest Cover Type Data



HPStream (+), SPDC (×), CluStream (∗), FP(0) (□), FP (o), FP(B) (△)

Conclusions

- Approach for incremental clustering under population drift
- Estimate number of clusters
- Drift detection problem has a well defined form, but remains difficult to solve
- Limit revision to part of clustering result that is invalid
- Performance depends on stability

Future Work

- Find low density hyperplanes directly (locally optimal solutions)
- Density free mode/anti-mode seeking
- Different data types (time series data)

References


Aggarwal, C. C., Han, J., Wang, J., Yu, P. S. (2003). A Framework for clustering evolving data streams.
In *Proceedings of the 29th international conference on Very large data bases*, Vol. 29, pages 81-92.


Hartigan, J. A. and Hartigan P. M. (1985). The dip test of unimodality.
In *The Annals of Statistics*, pages 70-84.


Tasoulis, S. K., Tasoulis, D. K., and Plagianakos, V. P. (2010). Enhancing principal direction divisive clustering.
In *Pattern Recognition*, Vol. 43, pages: 3391-3411.


Weng, J., Zhang, Y., and Hwang, W. (2003). Candid covariance-free incremental principal component analysis.
In *Pattern Analysis and Machine Intelligence*. Vol. 25, pages 1034-1040.