Unsupervised Assessment of Landscape Shifts Based on Persistent Entropy and Topological Preservation

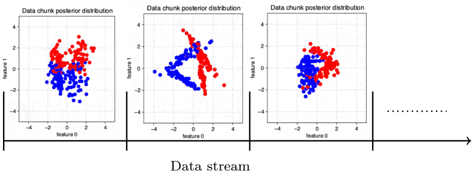
Sebastián Basterrech

DTU-Compute, Technical University of Denmark, Kongens Lyngby, Denmark

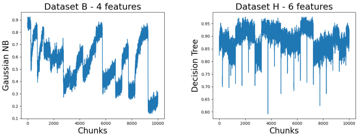
KDD'2024 – DELTA
August 26, 2024, Barcelona, Spain

Machine learning on streaming data
Dealing with shifts - common approaches

- How to handle the concept drift?



- Based in the predictive accuracy of the classifiers.
  - Only for supervised cases.
  - Latency problem: delay in obtaining ground truth labels.



- Based directly on the raw data.
  - Estimation of the probability mass function (pmf) is still a hard task.
  - Monitoring aggregation metrics (e.g. moving Average, CumSum) and statistical tests (e.g. KS).
  - Complex problem when the data belong to a high dimensional space, time limitation, scalability, etc..

- Emphasis: changes in the geometry of the patterns and/or changes in the data distribution.

Figures taken from: Ksieniewicz and P. Zyblewski, doi: 10.1016/j.neucom.2021.10.120
S. Basterrech and M. Wozniak, doi:10.1109/SMC53654.2022.9945547.

Proposed work

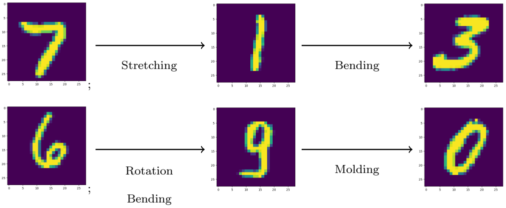
Sameness vs difference

- When two concepts are “essentially” the same?
- When two concepts are “essentially” different?

Proposed work

Sameness vs difference

- The “essence” of an object remains unchanged under simple continuous transformations (rotation, scaling, etc.)



- Equivalent objects in terms of topology (homeomorphic objects).

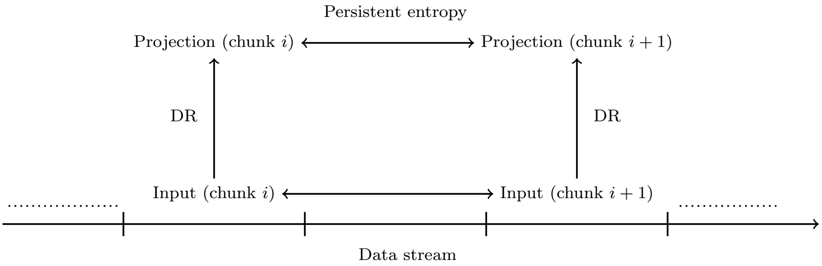
Roots:

- Inspired from: Spherical CNNs, developed by the team of Max Welling (ICLR’2018, Arxiv: 1801.10130).
- Initial study of topological preserved projections in: Basterrech, S., Clemmensen, L., Rubino, G.: A Self-Organizing Clustering System for Unsupervised Distribution Shift Detection. IJCNN’2024. Arxiv: 2404.16656.

Proposed work
Methodology

Highlights

- An attempt to broaden the concept drift paradigm.
- A novel approach to concept drift detection, leveraging algebraic topology and persistent entropy.
- Procedure: dimensionality reduction and persistent homology.
- Assessment: non-parametric statistical test.
- The method can be applied to both supervised and unsupervised contexts.



Methodology

Topology preserving clustering methods:

- "Essential" neighborhood relationships in the input space are preserved.

- Topology preserving property:

  If two inputs, $\mathbf{x}_1$ and $\mathbf{x}_2$, are "close" in the input space, then $\phi(\mathbf{x}_1)$ and $\phi(\mathbf{x}_2)$ are "close" in the projected space.
- Self-Organizing Maps (non-linear projection, specific NN).
- Baseline methods: PCA, Kernel-PCA.

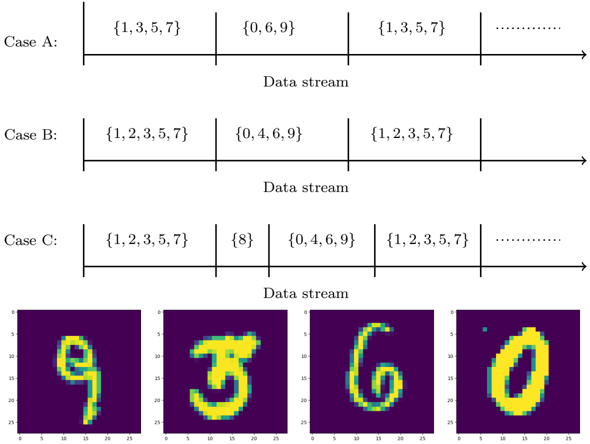
Persistent entropy

- We understand topological features as shapes that remain unchanged under certain continuous transformations.
- Persistent homology tracks changes in topological features of data across multiple scales.
- Persistent entropy provides a summary of the information derived from persistent homology (in only one scalar!).

Evaluation

How to make a proper evaluation?

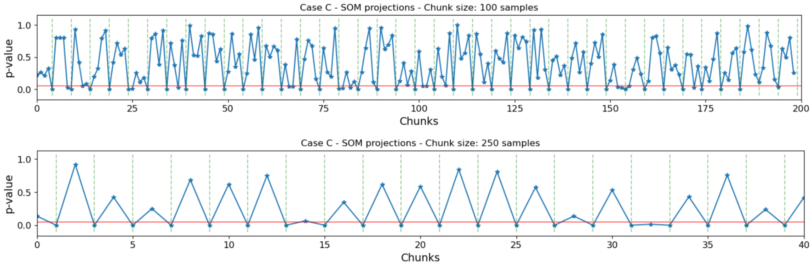
Figure: Generation of three categories of data streams.



Results

Evaluation

Example of results:



Summary:

- Framework provides a univariate signal with persistent entropy values.
- We apply a non-parametric statistical test (Mann-Whitney U test) for comparing consecutive chunks.

Discussion

Is it adequate to identify a drift if a sequence consists only of equivalent objects in terms of topology?

- Persistent homology as a tool for detecting significant changes between chunks of objects.
- Advantages: Usnsupervised, tracking changes using p-values, embedding topological information in a real sequence.
- Limitation: Initial experimental evaluation with promising results across synthetic data. However: hard to evaluate, missing annotated benchmark data.

Closing

Thank you!

Contacts:

- Email: sebbas@dtu.dk.
- ORCID: https://orcid.org/0000-0002-9172-0155

References

- G. Carlsson: Topology and Data. Bulletin of the American Mathematical Society (2), 255–308 (2009).
- P. Ksieniewicz, P. Zyblewski, Stream-learn-open-source Python library for difficult data stream batch analysis, Neurocomputing, 478, 2022, pp 11-21. Doi: 10.1016/j.neucom.2021.10.120.
- S. Basterrech and M. Woźniak, "Tracking changes using Kullback-Leibler divergence for the continual learning," SMC'2022, pp. 3279-3285, doi: 10.1109/SMC53654.2022.9945547.
- S. Basterrech, L. Clemmensen, G. Rubino: A Self-Organizing Clustering System for Unsupervised Distribution Shift Detection. IJCNN'2024. Arxiv: 2404.16656.