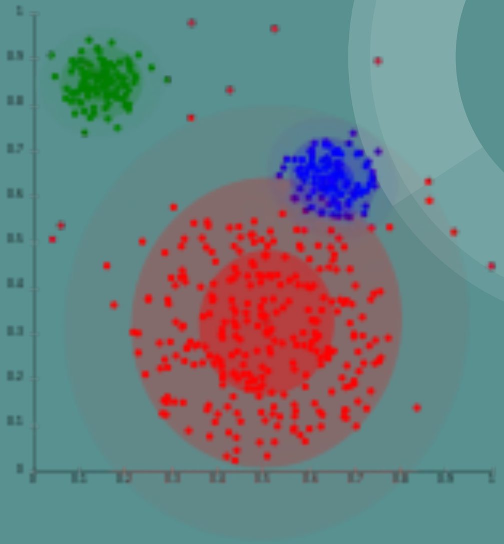
Dynamic Clustering



Alisa Leshchenko
Daniel Mallia
Vishnu Rampersaud

Hunter College CSCI 353/795 Machine Learning Fall 2020

Problem Description

“If intelligence was a cake, unsupervised learning would be the cake, supervised learning would be the icing on the cake, and reinforcement learning would be the cherry on the cake.” - Yann LeCun (Geron 235)

- Clustering - unsupervised classification - *dynamic* (no downtime)
- Online, incremental
- Concept drift - virtual (data) and real (target) - cluster evolution
- No guarantees about clusters - shape, densities
- Time dependent development
- Minimize parameters

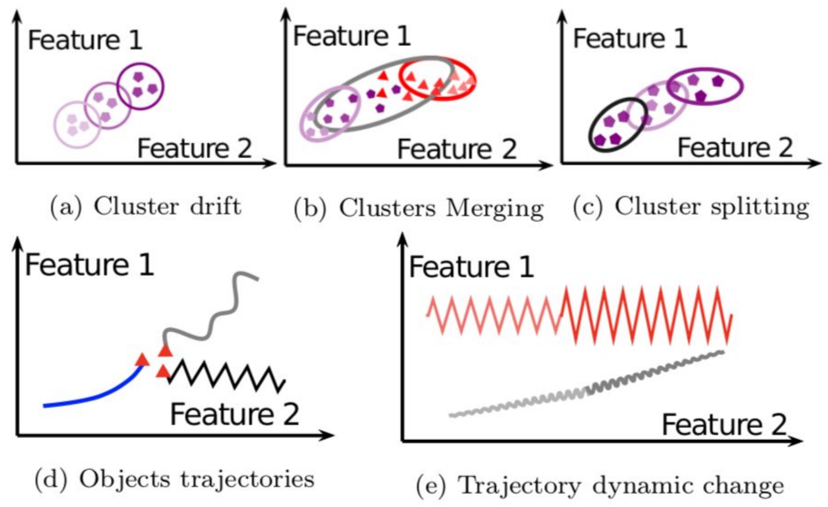
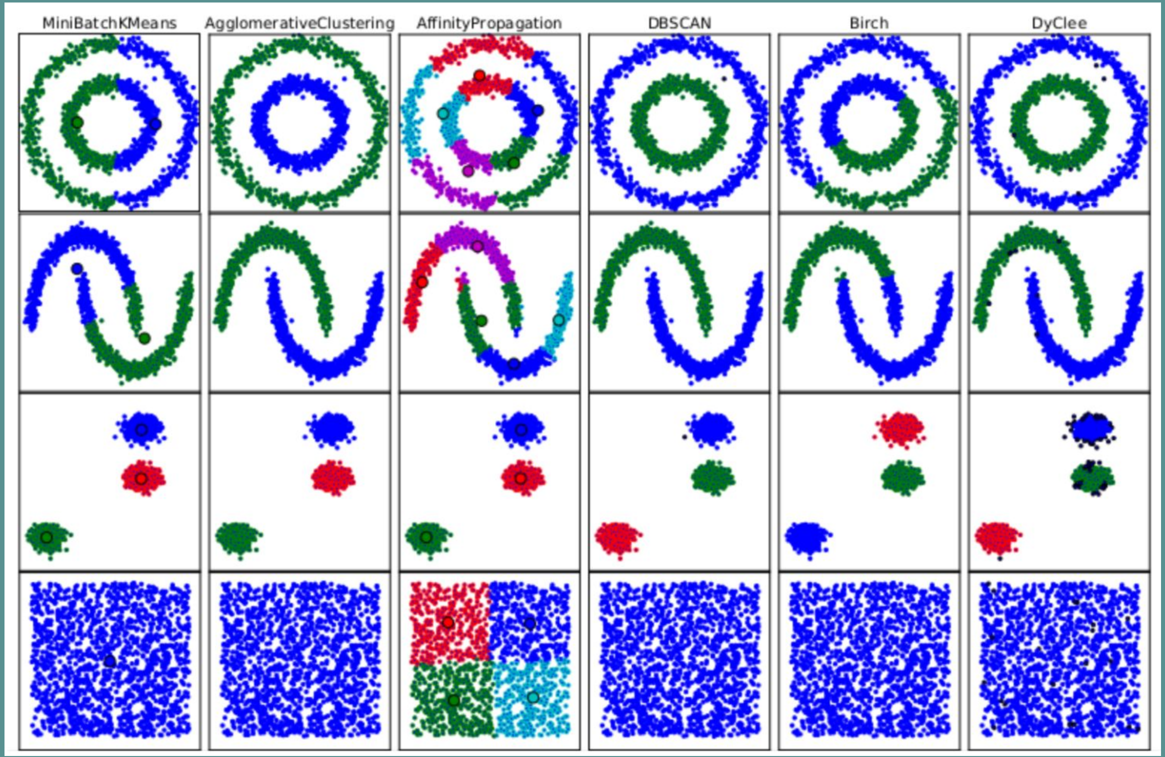


Fig. 1. Time-varying dynamic system’s data representation

[1] Page 1225

Dan



MiniBatchKMeans
AgglomerativeClustering
AffinityPropagation
DBSCAN
Birch
DyClee

[3] Page 19

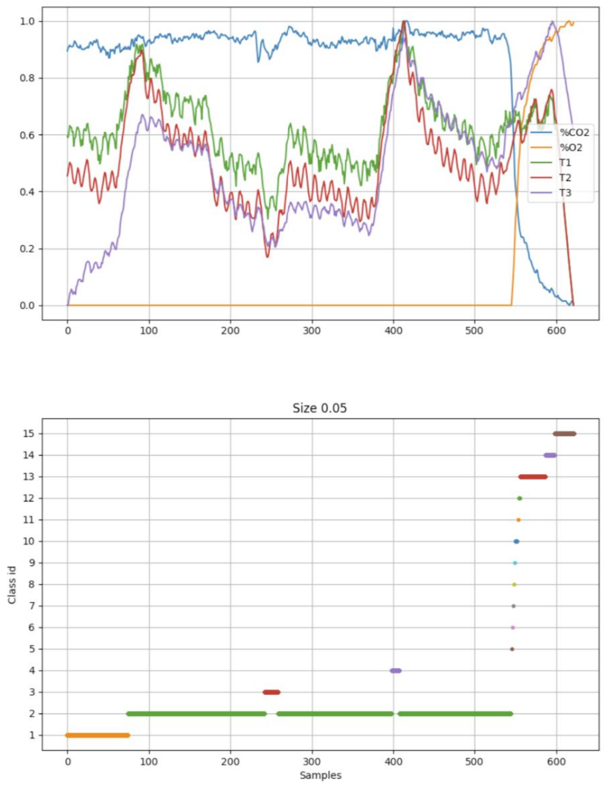


Figure 28: Gasifier normalized data scenario (top) and clustering from ***DyClee*** (bottom)

Cluster Id	Identification
Class 1	Normal operation in start up mode (temperature T3 in the reactor is not yet quite high)
Class 2	Normal operation
Class 3	Detection of combustion problem
Class 4	High steam flow state
Class 13	No solid feed state
Class 15	Shut down state
Classes 5 to 12 and 14	Transitory operating states

Table 10: Identification of the clusters provided by ***DyClee*** for the gasifier scenario

What is Dynamic Clustering and why is it interesting?

- A static distribution is not assumed — the clusters adapt to a changing environment
- Useful in many real-world applications
- Example: state of a machine over the course of operation (from powering on to powering off)
- Can be used to diagnose malfunctions

[3] Page 30

Alisa

Related Work

- Two stage approaches
- Distance/density
- Snapshots
- Indexing methods (ex: ClusTree)
- Stream speed handling
- Concept drift

[1] Nathalie Andrea Barbosa Roa, Louise Trav´e-Massuy`es, Victor Hugo Grisales Palacio. Trend-Based Dynamic Classification for on-line Diagnosis of Time-Varying Dynamic Systems. 9th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes, IFAC, Sep 2015, Paris, France. hal-01205313 https://hal.archives-ouvertes.fr/hal-01205313/document

[2] Nathalie Andrea Barbosa Roa, Louise Trav´e-Massuy`es, Victor Hugo Grisales Palacio. A novel algorithm for dynamic clustering: properties and performance 016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Dec 2016, Anaheim, United States. pp.565-570, 10.1109/ICMLA.2016.0099.hal-02004417 https://hal.archives-ouvertes.fr/ hal-02004417/document

[3] Nathalie Barbosa Roa, Louise Trav´e-Massuy`es, Victor Hugo Grisales. DyClee: Dynamic clustering for tracking evolving environments. Pattern Recognition, Elsevier, 2019, 94, pp.162-186. 10.1016/j.patcog.2019.05.024 . hal-02135580 https://hal.laas.fr/hal-02135580/document

[4] W. Shao, L. He, C. Lu, X. Wei and P. S. Yu, "Online Unsupervised Multi-view Feature Selection," 2016 IEEE 16th International Conference on Data Mining (ICDM), Barcelona, 2016, pp. 1203-1208, doi: 10.1109/ICDM.2016.0160.

Approach

- Implement the DyClee algorithm:
  - A fully Dynamic Clustering algorithm for tracking Evolving Environments, that handles non-convex, multi-density clustering with outlier rejection even in highly overlapping situations.
  - Dynamic classification where the classifier structure changes according to the input data
  - Uses online-offline (distance-density) stages clustering approach found in the CluStream and other algorithms - micro-clusters into final clusters
  - Different methods of indexing micro-clusters
- Novel aspects:
  - Online feature reduction
  - More informative density stage

Dan

Team Roles

Alisa - Primary focus on mathematical background. Secondary focus on LATEX formatting, poster, and general presentation. Tertiary focus on implementation and novel approaches.

Daniel - Primary focus on design and Python implementation of the DyClee algorithms. Secondary focus on efficiency improvements. Tertiary focus on visualizations.

Vishnu - Primary focus on design and Python implementation of the DyClee algorithms. Secondary focus on data visualizations. Tertiary focus on efficiency improvements.

Evaluation

- Replicate selected tests used by the authors to ensure we are achieving comparable results
    - Synthetic Scikit-learn datasets - static classification, path-based clustering
    - Toy example using synthetic data for dynamic classification
- Evaluation of novel component
    - Evaluate performance using the above results as benchmarks
- Real datasets/tasks:
    - Time-series dataset - CSTH, MIMIC, or UCI dataset

Alisa

Timeline

Week Of:	Deliverables:
Nov. 3	Milestone 2
Nov. 10	Scikit-Learn experiments with clustering and plotting; DyClee module and class interfaces
Nov. 17	Micro_cluster and Cluster_group classes with relevant functions
Nov. 24	DyClee - paper 1 (minus episode characterization) working
Dec. 1	Implementing paper 3 optional behaviors; testing automatic feature selection (and/or additional trend extraction)
Dec. 8	Finalized Code
Dec. 15	Final deliverables (Code, paper, poster, demo)

Vishnu