ML Final Project Proposal: Dynamic Clustering

Alisa Leshchenko[1], Daniel Mallia[2], and Vishnu Rampersaud [3]

1 Project Description

In the present day, the volume of data available regarding every facet of society and our world more generally is increasing, as is the speed at which it is being generated and shared. While supervised learning is an inherently limited task, requiring labeled data and assumptions regarding how representative training data will be of test data and real world application, unsupervised learning allows for knowledge discovery and distillation, and even classification in semi-supervised or active learning applications, for this avalanche of data. How to perform this unsupervised learning with minimal input from a user, for example not needing to specify how many clusters to expect, in an online context, with no assumptions about or guarantees for cluster shape and evolution, or even data input speed, is a challenging problem that is the subject of current and exciting research. Furthermore, the applicability of algorithms which satisfy these criteria is broad: industrial contexts with sensor streams, vitals monitoring in a medical context, or perhaps even video segmentation, identifying the presence of objects or behaviors for review amongst voluminous security camera data.

To that end, we propose an algorithmic project with a goal of implementing (from scratch) and improving upon the algorithms described across three papers that culminated in the DyClee algorithm described in the third paper "DyClee: Dynamic clustering for tracking evolving environments" cited below. This project falls under the broad category of unsupervised learning and specifically focuses on the task of clustering. As defined in the DyClee papers referenced below, the algorithm(s) on which we hope to improve encompass a clustering approach that is suited to a truly online context, where time-series data must be clustered in real-time, as in the case of sensor streams in industrial contexts. To that end, the approach also includes an emphasis on adaptability: being able to effectively cluster in the face of dynamic and changing data, without needing to go offline for an "overhaul" in restructuring or retraining.

To this, we hope that we might bring an emphasis on greater efficiency, both in terms of cluster searching and in parallel feature selection or dimensionality "reduction". The former would likely involve using final clusters to limit the number of distance calculations which must be performed in the first distance-based stage. The latter stems from interest in how online ranking and reevaluation of feature explanatory power could offer opportunities to perform lower dimensional calculations or, if the user chooses only a certain number of features amongst which micro-clusters must agree to be considered as connected, put emphasis on useful features. If time allows, we are interested in exploring the episode abstraction dynamics which allow for broader understanding of events and trends in sequential data. In a computer vision video segmentation context this could mean

[1] Undergraduate, Mathematics

[2] Graduate, Computer Science

[3] Graduate, Computer Science

identifying the arrival and departure of new objects in a video stream; in a medical context such as the MIMIC database, or even in the industrial contexts discussed in the DyClee papers, this can mean identifying worrying or even dangerous states. Dynamic clustering in particular is applicable in these contexts because they involve changing environments with high-volume data streams.

2 Team Member Contributions

Alisa's Contributions: Primary focus on mathematical background. Secondary focus on LaTeX formatting, poster, and general presentation. Tertiary focus on implementation and novel approaches.

Daniel's Contributions: Primary focus on design and Python implementation of the DyClee algorithms. Secondary focus on efficiency improvements. Tertiary focus on visualizations.

Vishnu's Contributions: Primary focus on design and Python implementation of the DyClee algorithms. Secondary focus on data visualizations. Tertiary focus on efficiency improvements.

3 CSCI 353/795 Related Topics

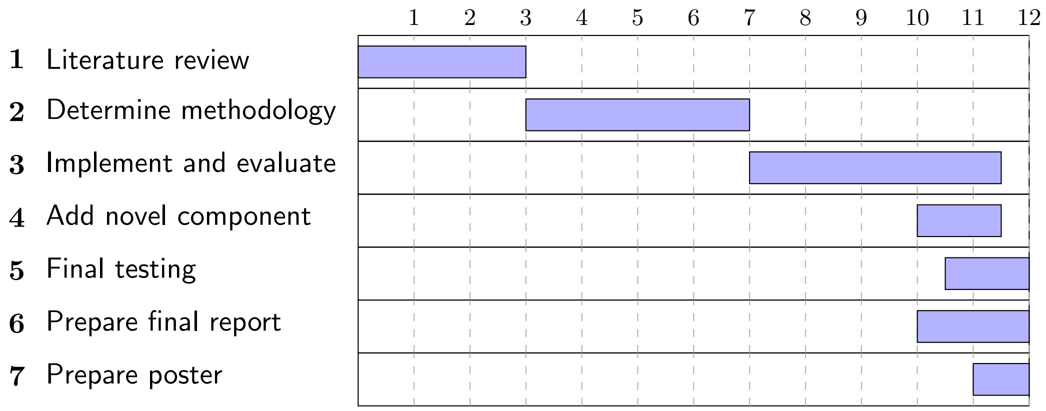
All groups members will be focused on the following aspects:

1. Unsupervised learning, particularly clustering and outlier detection, with possible usage of dimensionality reduction
2. Meaningful processing of data streams – sequential and time-series data – for understanding events and trends
3. Dealing with dynamic and evolving big data, where learning must be done incrementally in real-time without downtime and without guarantees of cluster convexity, separation or persistence

4 Datasets

Dataset	**Notes**
Synthetic Datasets and Common Benchmarks	Recreating original paper evaluations.
Time-Series Data	A small UCI dataset for testing dynamicity.
The Continuous Stirred Tank Heater Simulation	Dataset used in [1] - stretch goal evaluation.
MIMIC-II	Vitals time-series data - stretch goal evaluation.

5 Timeline (in weeks, beginning week of Sept 13th)



Reference: week 7 is the week of November 1st

Week Of	Deliverables
Nov. 3	Milestone 2
Nov. 10	Scikit-Learn experiments with clustering and plotting; DyClee module and class interfaces
Nov. 17	MicroCluster and ClusterGroup 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)

6 Deliverables

1. Minimal viable product: A full implementation of the algorithm as described in the 2015 paper, with novel features (cluster search / dimensionality reduction)

    i) Stretch goal: Apply the algorithm to MIMIC-II or other time-series data

2. A demo including a video recording of real-time clustering in progress

3. A 30-page final report detailing our methods and findings

4. A conference style poster

7 Project Evaluation

Evaluation of the project will proceed along two tracks: unlabelled classification results in terms of clusters formed and outliers detected, and runtime efficiency, as this approach is designed to be suitable for real-time applications without downtime. To that end, the former will involve a comparison of results to synthetic scikit learn datasets that were discussed in [3]. These tests will also serve as a benchmark to test the novelty component that will be introduced in this project. Time permitting, we will also evaluate the datasets found in [1] and test for successful identification of clusters or outliers as pre-identified for additional datasets, such as MIMIC-II. The latter will primarily involve our own process of benchmarking, as the cited papers only provide general complexity analysis.

8 References

[1] Nathalie Andrea Barbosa Roa, Louise Travé-Massuyès, 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 Travé-Massuyès, Victor Hugo Grisales Palacio. *A novel algorithm for dynamic clustering: propertiesand performance* 016 15th IEEE International Conference on Machine Learning andApplications (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 Travé-Massuyès, 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.