Dynamic Clustering: A Novel Take on DyClee

Alisa Leshchenko[1], Daniel Mallia[1], Vishnu Rampersaud[1]

[1]Department of Computer Science, Professor: ANITA RAJA

Introduction

DyClee is a dynamic clustering algorithm that works in two phases: a fast distance-based clustering stage that pre-processes data in order to create preliminary micro-clusters, and a slower density-based clustering stage that groups the proto-clusters into final clusters. The use of a pre-processing stage allows the algorithm to work with high-volume live data, as it provides a way of working with highly compressed summary data and does not necessitate storage or processing of the whole dataset. Density-based clustering allows the algorithm to tackle clusters of arbitrary shape, including non-convex clusters; it also allows the algorithm to deal with multi-density data, which is important for applications, where uniform density cannot generally be assumed.

In addition, DyClee is equipped with a forgetting process which gives it the ability to detect novel concepts, reject outliers, and be responsive to concept drift (the evolution of the data input distribution and/or the conditional distribution of the target variable). At the time of publication of reference paper 3, no other algorithm satisfactorily addressed all of the above capabilities.

Approach

Our goal was to meet a baseline of recreating the 2015 paper, possibly without episode abstraction, while including any helpful optional parameters and behaviors from the subsequent papers. This entailed a full implementation of the distance and global density stages, as well as decay functions for the forgetting process. We initially intended to expand on this baseline with work on feature selectivity and efficient usage of spatial search; the latter was dropped in favor of correcting issues the authors' approach exhibited with labeling clusters.

$$L_\infty(X, c_k) \equiv \max_i \left|x^i - c_k^i\right| < \frac{S^i}{2}, \forall i = 1, \ldots, d$$

Definition of a microcluster reachable from a data sample X (Barbosa Roa, 2019)

Experiments

DyClee was first tested on several synthetic datasets to benchmark it against the authors' performance. Among these were the Scikit-Learn make_circles, make_moons, and make_blobs dataset, and the Chameleon dataset. Next it was tested on simple image data, as well as on an urban road accident geospatial coordinates dataset.

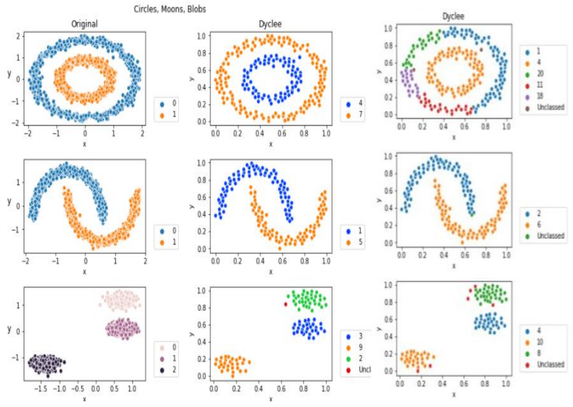
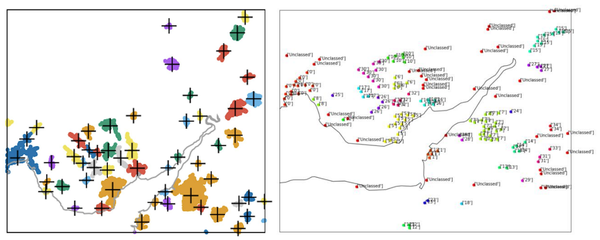


Figure 1

Figure 2



Part of the UrbanGB dataset, a dataset of urban road accident locations in Great Britain. The dataset poses multiple changes for a clustering algorithm, including the forms of the clusters and the vast number of them. Left is the results of the k-means variant proposed in the paper which presented the dataset's release. Right is our limited results. See paper for citation.

Novelty

- **Feature Selectivity**: In defining micro-cluster hyperbox overlap, the DyClee authors specify a parameter, varphi, which dictates the minimum number of dimensions in which the hyperboxes must overlap to be considered connected. However, no mechanism is specified for selecting which dimensions. To this we added an optional behavior which specifies that the varphi dimensions with the greatest observed variance thus far must be agreed upon.

- **Label Propagation**: The DyClee authors did not specify any order for processing connected micro-clusters or for handling labels in case of cluster merging. We propose a prioritization of labelled micro-clusters and a label selection scheme which preserves user input labels if present, and otherwise resolves label selection by mode. If there are multiple modes, micro-clusters vote by density and inverse distance to project final cluster center, to choose a label.

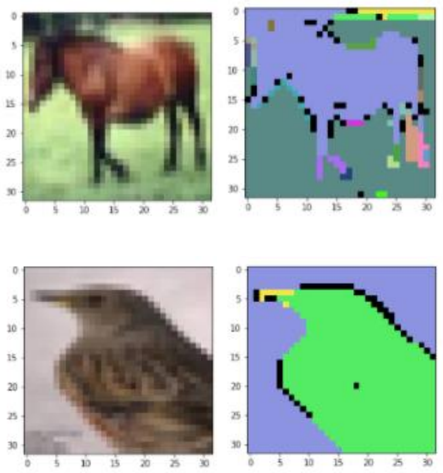


Figure 3: Segmentation results. See paper for dataset citation.

Results

Performance of Novelty Components: Figure 2 shows the performance of baseline DyClee on several synthetic datasets. Figure 1 illustrates the drastic performance improvements achieved by using our novel label propagation scheme.

Image data: DyClee performed rudimentary but fairly effective image segmentation on small images from the CIFAR10 dataset (Figure 3). Cluster colors were chosen randomly, with the exception of black marking outliers.

Urban traffic data: While we had to limit our testing to a small portion of the dataset (so the microclusters represent only a portion of the data), it can clearly be seen that DyClee offers comparable clustering, with outlier information and no requirement to specify the number of clusters.

Future Work

- **Efficiency improvements**: Parallelization and KD Tree for faster instance insertions
- **Episode extraction**: Wavelet analysis will be applied to allow for episode abstraction and characterization

Conclusions

DyClee is a simple and potent algorithm that achieves a number of desirable features in a clusterer; with a couple of innovations, it can be expanded to more reliable and robust performance. However, its seemingly simple phi parameter can be unintuitive in practice and the steep cost complexity of its operation can be prohibitive for all but the most efficient lower-level language implementations.

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

2. Nathalie Barbosa Roa, Louise Travé-Massuyès, Victor Grisales. A novel algorithm for dynamic cluster- ing: properties and performance. 2016 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

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.