

A DENSITY-GATED COVERAGE SAMPLER FOR CROSS-DOMAIN ACTIVE LEARNING IN BIOACOUSTICS

Technical Report

Ragib Amin Nihal, Benjamin Yen, Takeshi Ashizawa, Kazuhiro Nakadai

Institute of Science Tokyo
Department of Systems and Control Engineering
Tokyo, Japan

1. ABSTRACT

We present `capped_rarity_kcenter`, an acquisition function for BioDCASE 2026 Task 4 that operates on frozen Perch v2 embeddings and requires no per-dataset tuning. The backbone is a greedy k-center (farthest-first) over a cached PCA-128 projection of the embeddings, which provides the spread-coverage needed to reach minority-class regions that drive macro-mAP. On top of coverage we layer three terms: (i) a *capped class-deficit rarity* term that draws toward under-labelled classes but stops once a class reaches its balanced share (the cap is essential to avoid destabilising near-single-label subsets), (ii) a *budget-ramped mean-margin uncertainty* term computed from a kNN label surrogate, and (iii) a *metadata-only location-stratification* term that compensates for recording-site structure surviving into the embeddings. A single density gate, keyed to the average number of labels per sample, fades these terms in or out: sparse subsets default to near-pure coverage, while denser ones receive the full signal. This is what allows one configuration to serve both the Antarctic whale data (ATBFL) and the avian data (HSN/POW/UHH). A matching adaptive warmup handles the first cycle on the same location signal. The method is deterministic given a seed, uses only revealed labels, and runs in well under a second per cycle on a single CPU. Averaged over the four development subsets (5 repeats, 500-label budget), it reaches an AULC of **0.4825**, above the CoreSet baseline (0.4728), at a lower sampling cost.