Efficient Spam Detection via Class-Balanced Uncertainty-Density Ranking Coresets

Motivation. Spam detection in online communication remains challenging due to severe class imbalance, rapidly evolving spam patterns, and the high computational demands of Transformer-based models. Supervised learning approaches often underperform when spam messages are underrepresented, while training on full datasets incurs significant cost. Existing coreset selection methods typically focus on either uncertainty-based sampling (e.g., entropy, margin, confidence) or diversity-based selection (e.g., clustering, k-center, representativeness). However, these strategies often overlook class imbalance and fail to jointly capture both informativeness and representativeness. This highlights the need for efficient data reduction techniques that maintain predictive accuracy while reducing annotation and training overhead.

Method. We propose a novel coreset selection framework, *Class-Balanced Uncertainty-Density Ranking (CBUDR)*, which simultaneously captures predictive uncertainty, representativeness, and class balance. Each sample is assigned a class-normalized uncertainty score, $U_c(x_i) = \frac{U(x_i)}{\max_{x \in C_c} U(x)}$, mitigating over-prioritization of minority or noisy samples. To ensure geometric coverage, a density score $D(x_i) = 1 - \frac{1}{|N_i|} \sum_{x_j \in N_i} \text{sim}(e_i, e_j)$ is computed, where N_i denotes the k-nearest neighbors in embedding space and sim is cosine similarity; higher scores highlight sparsely populated regions. These components are combined via a convex score, $\text{CBUDR}(x_i) = \alpha \cdot U_c(x_i) + \beta \cdot D(x_i)$ with $\alpha + \beta = 1$, providing a controlled trade-off between exploration (uncertainty) and coverage (representativeness). Samples are then ranked by this score, and the highest-ranked subset is selected as the coreset for training.

Results. We evaluate our approach, CBUDR, against random sampling and conventional uncertainty/diversity strategies across three benchmark datasets for SMS, email, and Twitter spam detection (Table 1). The results consistently show that class-wise Bottom-K with CBUDR achieves near-perfect accuracy and F1-scores ($\geq 99\%$) using only 5% of the training data, outperforming both random selection and Top-K uncertainty methods. On UtkMl Twitter and LingSpam, CBUDR not only surpasses the full-data baseline but also demonstrates that "easy yet representative" samples selected via Bottom-K ranking yield stronger generalization than traditional high-uncertainty examples, highlighting a previously underexplored regime of coreset design. These findings demonstrate that principled data selection can simultaneously improve efficiency and generalization.

Impact. CBUDR enables lightweight spam filtering systems that are computationally efficient and suitable for real-time or resource-constrained environments. By explicitly incorporating class balance, the method promotes equitable treatment of minority spam messages, which are often overlooked in standard selection criteria. Beyond spam detection, CBUDR provides a general framework for uncertainty-aware data reduction, with potential applications in fraud detection, misinformation filtering, and other domains requiring robust learning under imbalance.

Table 1: Performance of Different Coreset Selection Strategies and Ranking Methods on different datasets.

Dataset	Coreset Strategy	Ranking Method	5%				10%				25%			
			Acc (%)	F1 (%)	Prec (%)	Rec (%)	Acc (%)	F1 (%)	Prec (%)	Rec (%)	Acc (%)	F1 (%)	Prec (%)	Rec (%)
UtkMl Twitter	Random		94.44	93.98	**100.00**	88.64	94.44	94.12	97.56	90.91	95.55	95.41	98.11	92.86
	Class-wise Top-K	Entropy	63.33	67.33	59.65	77.27	83.33	81.01	90.14	73.56	90.42	90.02	91.08	88.99
		CBUDR	73.33	68.42	81.25	59.09	89.44	89.14	88.64	89.66	88.20	88.40	84.52	92.66
		Entropy+CBUDR	78.89	76.54	83.78	70.45	82.22	78.67	93.65	67.82	89.76	89.55	88.74	90.37
	Class-wise Bottom-K	Entropy	98.89	*98.85*	**100.00**	97.73	*98.33*	98.27	*98.84*	97.70	98.44	98.38	99.07	97.71
		CBUDR	**100.00**	**100.00**	**100.00**	**100.00**	**99.44**	**99.42**	**100.00**	98.85	**99.33**	99.31	**100.00**	98.62
		Entropy+CBUDR	98.89	98.88	97.78	**100.00**	*98.33*	98.25	**100.00**	96.55	*98.66*	98.61	99.53	97.71
	Baseline	All (100%)									96.49	96.41	95.92	96.91
UCI	Random	None	**100.00**	**100.00**	**100.00**	**100.00**	97.62	91.67	84.62	**100.00**	98.09	93.10	*90.00*	96.43
	Class-wise Top-K	Entropy	*90.48*	*60.00*	*60.00*	*60.00*	90.48	63.64	*63.64*	63.64	*99.04*	*96.30*	**100.00**	*92.86*
		CBUDR	*90.48*	33.33	**100.00**	20.00	*91.67*	*74.07*	62.50	90.91	97.61	90.91	92.59	89.29
		Combined	*90.48*	33.33	**100.00**	20.00	90.48	69.23	60.00	*81.82*	97.13	88.00	**100.00**	78.57
	Class-wise Bottom-K	Entropy	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	99.52	98.18	**100.00**	96.43
		CBUDR	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
		Combined	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	99.52	98.18	**100.00**	96.43
	Baseline	All (100%)									99.52	98.18	**100.00**	96.43
LingSpam	Random	None	90.91	50.00	**100.00**	33.33	88.64	*70.59*	*60.00*	85.71	*99.08*	*97.30*	*94.74*	**100.00**
	Class-wise Top-K	Entropy	*86.36*	*0.00*	0.00	*0.00*	*81.82*	55.56	45.45	*71.43*	87.16	61.11	61.11	61.11
		CBUDR	90.91	50.00	**100.00**	33.33	79.55	40.00	37.50	42.86	93.58	78.79	86.67	*72.22*
		Combined	81.82	*0.00*	0.00	*0.00*	77.27	37.50	33.33	42.86	95.41	83.87	**100.00**	*72.22*
	Class-wise Bottom-K	Entropy	**100.00**	**100.00**	**100.00**	**100.00**	**97.73**	**93.33**	87.50	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
		CBUDR	**100.00**	**100.00**	**100.00**	**100.00**	**97.73**	92.31	**100.00**	85.71	**100.00**	**100.00**	**100.00**	**100.00**
		Combined	**100.00**	**100.00**	**100.00**	**100.00**	**97.73**	92.31	**100.00**	85.71	**100.00**	**100.00**	**100.00**	**100.00**
	Baseline	All (100%)									99.54	98.61	98.61	98.61