Subclass-Aware Inclusive Classifier via Repulsive Hidden Strata

Anonymous Authors

Classification models in machine learning are typically trained with coarse-grained class labels, which overlook fine-grained subclass variations. This phenomenon, known as hidden stratification [1], results in asymmetric performance; models excel on dominant subclasses but struggle on rare or underrepresented ones. Such biases critically undermine fairness and robustness, especially in safety-sensitive applications such as medical imaging. We introduce **Subclass-Aware Inclusive Classification (SAIC)**, a framework shown in Figure 1 that explicitly addresses hidden stratification. SAIC operates in two stages: (i) *unsupervised subclass identification* using a repulsive point process (k-DPP [2]) to uncover diverse and representative latent subclasses without prior assumptions, and (ii) *subclass-aware classification* with Group Distributionally Robust Optimization (GDRO), which emphasizes minimizing worst-case subclass loss. Extensive experiments on four benchmark datasets (MNIST, CIFAR-10, Waterbirds, and CelebA) show that SAIC consistently improves robustness without compromising overall accuracy. Specifically, we compare against K-means- and GMM-generated subclasses [3, 1] and also give the accuracy obtained using true subclass labels, as given in Table 1. Beyond overall accuracy, SAIC's clustering module demonstrates superior subclass identification, closely matching true subclass counts, preserving rare subclass purity, and maintaining moderate runtime efficiency. SAIC provides a scalable solution to hidden stratification by combining diversity-aware subclass discovery with robust optimization, thereby enhancing fairness and reliability in high-stakes classification tasks.

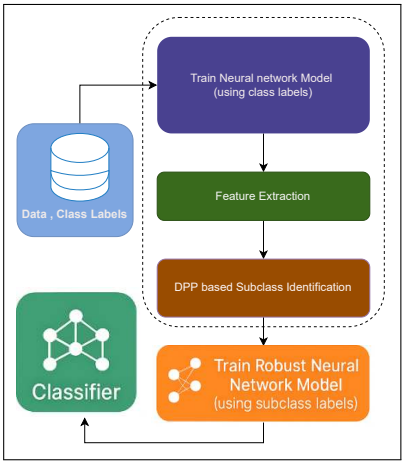

Figure 1: Flowchart of the proposed SAIC framework.

Method	MNIST	CIFAR 10	Water birds	CelebA
ERM	97.1	92.7	32.1	35.2
K-means	97.3	92.5	61.8	48.9
GMM	96.3	91.1	53.4	50.8
SAIC(ours)	**97.6**	**94.0**	**76.5**	**62.5**
TrueSubclass	97.0	93.1	87.8	85.2

Table 1: Summary of robust accuracy (%) across datasets.

References:

[1] N. Sohoni, J. Dunnmon, G. Angus, A. Gu, and C. Ré, "No subclass left behind: Fine-grained robustness in coarse-grained classification problems," Advances in Neural Information Processing Systems, 2020.

[2] A. Kulesza and B. Taskar, "k-dpps: Fixed-size determinantal point processes," in Proceedings of the 28th ICML, 2011.

[3] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967.