Fairness-Aware Class Imbalanced Learning on Multiple Subgroups

Davoud Ataee Tarzanagh*, Bojian Hou*, Boning Tong*, Qi Long, Li Shen
University of Pennsylvania

Label-imbalanced classification (LIC)

- LIC suffers from a significant discrepancy in the number of examples across classes, requiring the use of balanced accuracy as a more suitable metric than conventional misclassification error.
- Vector Scaling (VS) [2]:

$$\ell(f, \mathbf{v}; \mathbf{x}, y) = -u_y \log \frac{e^{\gamma_y f(\mathbf{x})_y + \Delta_y}}{\sum_{j=1}^{k} e^{\gamma_j f(\mathbf{x})_j + \Delta_j}}. \qquad \text{(VS)}$$

  Here, $\mathrm{v} = [v_1, \ldots, v_k]$, where $v_j = (u_j, \Delta_j, \gamma_j)$ are some hyperparameters to adjust the loss.
- Nonetheless, there is still a risk of overfitting on minority class samples despite these advancements [4].

Group-sensitive classification (GSC)

- In GSC, the goal is to ensure fairness concerning protected attributes like gender or race.
- Group Sufficiency (GS): We say that f is sufficient with respect to attribute A if $\mathrm{E}[Y|f(X)] = \mathrm{E}[Y|f(x), A]$.
- In overparameterized models with limited samples per subgroup, this control of group sufficiency may not always hold.

Fairness-Aware Class Imbalanced Learning on Multiple Subgroups (FACIMS)

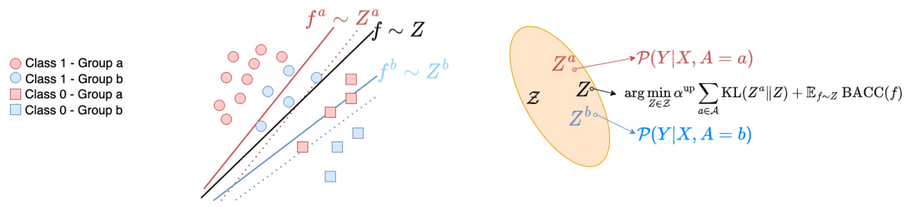


Figure 1: An illustration of the FACIMS model defined in (7). f^a and f^b maximize the margin for minority classes for groups a and b in (7b). In the upper level problem (7a), FACIMS finds $\mathbf{Z} \in \mathcal{Z}$ to achieve a small balanced accuracy while minimizing the discrepancy between $(\mathbf{Z}^{a,\star}, \mathbf{Z}^{b,\star})$. The approximation term $\mathrm{KL}(\mathbf{Z}^{a,\star}|\mathbf{Z})$ is based on the distribution family $\mathcal{Z}$ (orange region). If the predefined $\mathcal{Z}$ has good expressive power, the approximation is treated as a small constant.

FACIMS

$$\min_{\mathbf{v}, \mathbf{Z} \sim \mathcal{Z}} \quad \sum_{a \in \mathcal{A}} \alpha^{\mathrm{up}} \mathrm{KL}\left(\mathbf{Z}^{a,*} \| \mathbf{Z}\right) + \mathbb{E}_{\mathbf{w} \sim \mathbf{Z}}\, \mathcal{L}_{\mathrm{bal}}(\tilde{f}_{\mathbf{w}}; \mathcal{V}^a), \tag{7a}$$

$$\text{s.t.} \quad \mathbf{Z}^{a,*} \in \arg\min_{\mathbf{Z}^a \in \mathcal{Z}} \max_{\|\boldsymbol{\epsilon}^a\| \le \beta^a} \alpha^{\mathrm{low}}\, \mathrm{KL}(\mathbf{Z}^a \| \mathbf{Z}) + \mathbb{E}_{\mathbf{w}^a \sim \mathbf{Z}^a}\, \mathcal{L}_{\mathrm{vs}}(\tilde{f}_{\mathbf{w}^a + \boldsymbol{\epsilon}^a}, \mathbf{v}; \mathcal{T}^a), \quad \forall a \in \mathcal{A}. \tag{7b}$$

- FACIMS is a Bayesian-based tri-level optimization framework.
- In FACIMS, local predictors are learned using a small amount of training data and a fair, class-balanced predictor.
- The lower-level formulation utilizes the sharpness-aware minimization [1] to encourage convergence to a flat minimum and effectively avoid saddle points for minority classes.
- The upper-level problem dynamically adjusts the loss function by monitoring the validation loss, following a similar approach to [3], and updates the global predictor to align with all subgroup-specific predictors.

Theory

Assumptions

- $f_i(\boldsymbol{z}), \nabla f_i(\boldsymbol{z}), \nabla g_i(\boldsymbol{z}), \nabla^2 g_i(\boldsymbol{z})$ are $\ell_{f,0}, \ell_{f,1}, \ell_{g,1}, \ell_{g,2}$-Lipschitz continuous, respectively.
- $\nabla f_i(\boldsymbol{z}; \xi)$, $\nabla g_i(\boldsymbol{z}; \zeta)$, $\nabla^2 g_i(\boldsymbol{z}; \zeta)$ are unbiased estimators of $\nabla f_i(\boldsymbol{z})$, $\nabla g_i(\boldsymbol{z})$, $\nabla^2 g_i(\boldsymbol{z})$; and their variances are bounded.

Theorem

Under the above assumptions, if we choose the stepsize properly, then

- Convergence Rate $\le \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$
- Generalization Bound $\le \mathcal{O}\left(\left(\frac{p \ln A_\beta + \ln \frac{1}{\delta} + \ln(n|\mathcal{A}|)}{n|\mathcal{A}|}\right)^{\frac{1}{2}}\right)$.

Conclusion

- We studied fairness-aware class imbalanced learning on multiple subgroups (FACIMS) using a Bayesian-based optimization framework.
- Through extensive empirical and theoretical analysis, we demonstrated that FACIMS enhances the generalization performance of overparameterized models when dealing with limited samples per subgroup.

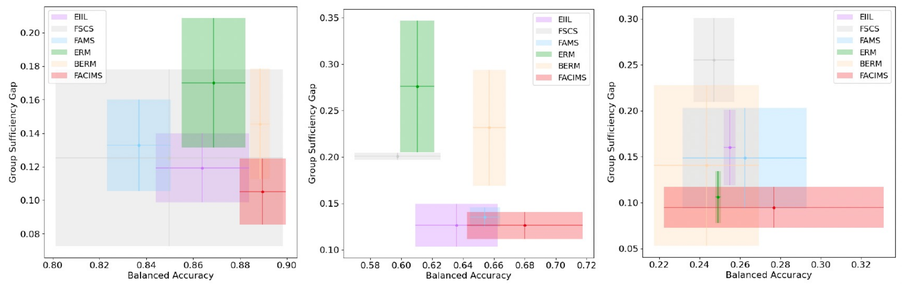
Experiments

Table 1: Statistical summary of the datasets including class and sensitive feature information.

Dataset	#Instance	#Features	Class	Class Distr.	Sensitive Feature	Sensitive Feature Distr.
Alzheimer's Disease	5137	17	AD / MCI	21% / 79%	Race	93.75% / 3.20% / 1.88% / 1.17%
Credit Card	30,000	22	Credible / Not Credible	22% / 77%	Education Level	46.77% / 35.28% / 16.39% / 0.93% / 0.41% / 0.17% / 0.05%
Drug Consumption	1885	9	Never used / Not used in the past year / Used in the past year / Used in the past day	1.81% / 5.41% / 65.98% / 26.80%	Education Level	6.74% / 6.90% / 26.86% / 14.28% / 25.48% / 15.02% / 4.72%

Table 2: Numerical results (mean ± standard deviation) for 5 repeats of different methods on Alzheimer's disease (AD) and Credit Card (CC) datasets regarding six measurements. Time is in the format of "hours:minutes:seconds". FACIMS-I means FACIMS ($\beta = 0$), and FACIMS-II means FACIMS ($\beta = 0$, $v = \bar{v}$). "↑" indicates the larger the better while "↓" indicates the smaller the better. The best one in each column is bold.

Data	Method	Balanced Accuracy ↑	Demographic Parity ↓	Equalized Odds ↓	Sufficiency Gap ↓	Recall 0 ↑	Recall 1 ↑	Time ↓
AD	EIIL	.8639±.0199	.0764±.0176	.1015±.0529	.1193±.0206	.9288±.0119	.7991±.0409	0:03:32
	FSCS	.8498±.0485	.0711±.0287	.1650±.1008	.1254±.0528	.9504±.0426	.7493±.1018	0:08:05
	FAMS	.8369±.0136	**.0431±.0210**	.1444±.0435	.1328±.0273	.7624±.0077	.9114±.0096	0:09:51
	ERM	.8687±.0136	.0550±.0196	.1143±.0390	.1701±.0387	**.9883±.0053**	.7491±.0430	**0:00:51**
	BERM	.8886±.0042	.0869±.0204	.0813±.0129	.1456±.0330	.9854±.0043	.7918±.0520	0:02:24
	FACIMS-II	.8839±.0079	.0747±.0182	.0868±.0130	.1167±.0139	.8456±.0148	**.9222±.0043**	0:09:58
	FACIMS-I	.8887±.0066	.0893±.0080	**.0450±.0049**	.1059±.0060	.8780±.0104	.8994±.0148	0:13:38
	FACIMS	**.8897±.0098**	.0765±.0208	.0616±.0142	**.1052±.0197**	.8832±.0072	.8962±.0054	0:15:26
CC	EIIL	.6357±.0267	.0834±.0200	.1723±.0515	.1266±.023	.7897±.0176	.4817±.0448	0:03:30
	FSCS	.5976±.0277	.0850±.0137	.2000±.0456	.2007±.0039	.8953±.0130	.3000±.0685	0:42:10
	FAMS	.6542±.0098	.0746±.0066	.1859±.0368	.1352±.0106	.8194±.0374	.4890±.0270	0:10:21
	ERM	.6104±.0111	.0599±.0173	**.1577±.0175**	.2760±.0710	**.9919±.0233**	.2289±.0820	**0:02:07**
	BERM	.6570±.0106	.1060±.0125	.1631±.0304	.2315±.0623	.8717±.0191	.4423±.0146	0:02:09
	FACIMS-II	.6446±.0163	.0707±.0073	.1973±.0358	.1340±.0147	.8002±.0374	.4890±.0270	0:10:03
	FACIMS-I	.6768±.0040	.0750±.0105	.1951±.0524	.1396±.0081	.8081±.0114	.5455±.0098	0:14:07
	FACIMS	**.6799±.0374**	**.0593±.0070**	.1567±.0230	**.1264±.0145**	.8136±.0054	**.5462±.0017**	0:14:18



(a) Alzheimer's Disease (b) Credit Card (c) Drug Consumption

Figure 2: Boxplot comparing balanced accuracy and group sufficiency gap for three real datasets with 5 repeats. The mean is represented by the middle of each box, while the box width represents twice the standard deviation. Better performance is indicated by boxes located towards the bottom right (higher balanced accuracy and lower group sufficiency). Two FACIMS variants are excluded for clarity, with complete results available in the appendix.

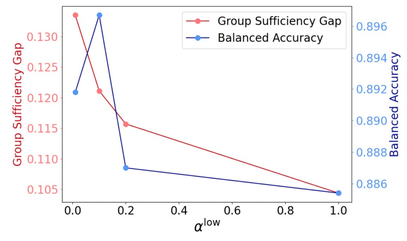


Figure 3: Accuracy-$\mathbf{SGap}_f$ curve under different α^{low} in Alzheimer's disease dataset.

In the middle level, the parameter α^{low} determines the attention given to $\mathrm{KL}(Z^a|Z)$. A higher value of αlow brings the local model closer to the global model, leading to improved group sufficiency gap but potentially worse balanced accuracy. We experimented with four different values of α^{low}: 0.01, 0.1, 0.2, and 1. Figure 3 illustrates the Accuracy-$\mathbf{SGap}_f$ curve under varying α^{low} on the Alzheimer's disease dataset. The figure demonstrates a clear trend: as α^{low} increases, both the balanced accuracy and group sufficiency gap decrease, aligning with our expectations.

Acknowledgements

This work was supported in part by the NIH grants U01 AG066833, RF1 AG063481, U01 AG068057, R01 LM013463, P30 AG073105, and U01 CA274576, and the NSF grant IIS 1837964. Data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative database (adni.loni.usc.edu), which was funded by NIH U01 AG024904. The authors Davoud Ataee Tarzanagh, Bojian Hou and Boning Tong have contributed equally to this paper.

Reference

[1] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur. Sharpness-aware minimization for efficiently improving generalization. arXiv preprint arXiv:2010.01412, 2020.
[2] G. R. Kini, O. Paraskevas, S. Oymak, and C. Thrampoulidis. Label imbalanced and group-sensitive classification under overparameterization. Advances in Neural Information Processing Systems, 34:18970–18983, 2021.
[3] M. Li, X. Zhang, C. Thrampoulidis, J. Chen, and S. Oymak. Autobalance: Optimized loss functions for imbalanced data. Advances in Neural Information Processing Systems, 34:3163–3177, 2021.
[4] H. Rangwani, S. K. Aithal, M. Mishra, and R. V. Babu. Escaping saddle points for effective generalization on classimbalanced data. arXiv preprint arXiv:2212.13827, 2022.