Measuring Bias and Fairness in Multiclass Classification

Cody Blakeney
Department of Computer Science
Texas State University
cjb92@txstate.edu

Gentry Atkinson
Department of Computer Science
Texas State University
gma23@txstate.edu

Nathaniel Huish
Department of Computer Science
Texas State University
njh71@txstate.edu

Yan Yan
Department of Computer Science
Texas State University
yyan34@iit.edu

Vangelis Metsis
Department of Computer Science
Texas State University
vmetsis@txstate.edu

Ziliang Zong
Department of Computer Science
Texas State University
ziliang@txstate.edu

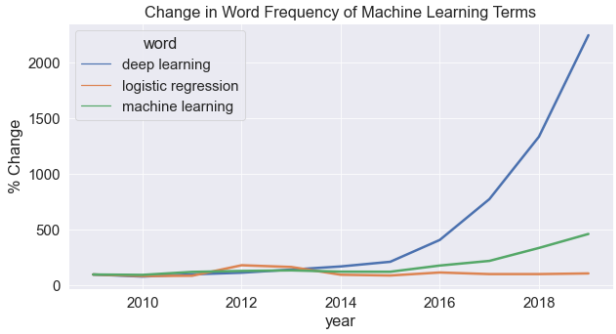
***Abstract*—Algorithmic bias is of increasing concern, both to the research community, and society at large. Bias in AI is more abstract and unintuitive than traditional forms of discrimination and can be more difficult to detect and mitigate. A clear gap exists in the current literature on evaluating the relative bias in the performance of multi-class classifiers. In this work, we propose two simple yet effective metrics, Combined Error Variance (CEV) and Symmetric Distance Error (SDE), to quantitatively evaluate the class-wise bias of two models in comparison to one another. By evaluating the performance of these new metrics and by demonstrating their practical application, we show that they can be used to measure fairness as well as bias. These demonstrations show that our metrics can address specific needs for measuring bias in multi-class classification. Demonstration code is available at https://github.com/gentry-atkinson/CEV_SDE_demo.git.**

***Index Terms*—fairness, bias, model quality, model comparison**

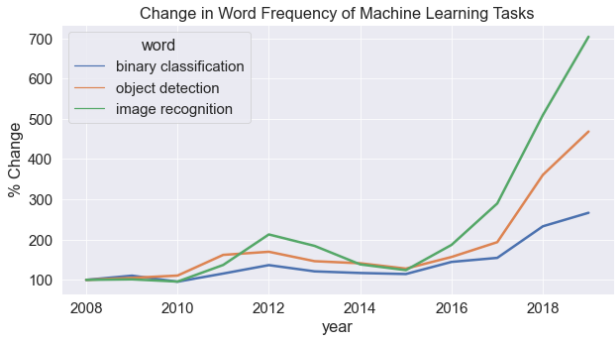
I. Introduction

Broad acceptance of the large-scale deployment of AI and neural networks depends on the models' perceived trustworthiness and fairness. However, research on evaluating and mitigating bias for neural networks in general and compressed neural networks in particular is still in its infancy. Because deep neural networks (DNNs) are "black box" learners, it can be difficult to understand what correlations they have learned from their training data, and how that affects the downstream decisions that are made in the real world. Two models may appear to have very similar performance when only measured in terms of accuracy, precision, etc. but deeper analysis can show uneven performance across many classes. Moreover, when the number of tasks grows beyond one or two, the difficulty in reasoning and quantifying trade-offs when selecting or validating a model also grows.

Widely accepted and effective metrics for measuring the bias of several neural networks against one another are still missing. Issues of both fairness and bias, which will be discussed as distinct but related phenomena in this paper, can seriously degrade the trustworthiness of a machine learning model in real-world conditions. It is important to quantify the performance of models in terms both of bias and fairness. While there exists extensive work on AI fairness regarding binary classification tasks [2]–[5], there is a shortage of metrics extending these ideas to other machine learning tasks.



(a) Deep Learning vs Logistic Regression



(b) Change in Task Word Frequency

Fig. 1: Google NGram Data [1] showing relative usage of machine learning related terms over time. Deep learning has quickly passed up the use of statistical terms like logistic regression. Computer Vision tasks like Object Detection and Image Recognition are growing at faster rates than Binary Classification which fairness metrics can address.

Even though many researchers are interested in studying the fairness of their models, we simply do not have the tools to measure for many domains yet.

In fact, recent trends shown in Figure 1, are exacerbating the divide with the majority of new research in neural networks exploring topics outside of the reach of existing fairness and bias metrics. While difficult to quantify exactly we see from Google NGram data [1] since 2010, the focus of machine learning is increasingly on multi-class classification and other difficult to quantify tasks rather than binary classification, revealing a need for metrics that can accommodate multi-class tasks. Worse still, we appear to be at the edge of another inflection point in AI where Large Language Models (LLMs) and Foundational models [6] like GPT-3 [7] are ingesting the entire corpus of human thoughts with limited supervision.

In this paper, we introduce two new metrics based on simple principles, whose purpose is to quantify a change in per-class bias between two or more models. These metrics provide singular data points that are easier to consider than the laborious checking of distributions of class-wise error rates. We will discuss their intuition and their application for comparing the relative performance of deep learning models, and classifiers in general, in terms of bias and fairness. To the best of our knowledge, these new metrics are distinct from all existing methods in that they expose per-group, per-class bias not neatly captured by other metrics, enabling the examination of issues of fairness and bias in great depth. While our proposed work is not a panacea for all emerging trends in AI, we believe it represents a starting point to address the current gap.

The remaining sections are organized as follows. In Section II we will contextualize the field of fairness metrics and their shortcomings as they relate to our considered problem domain. In Section III we will define the intuition for our metrics, and provide a mathematical definition. In Section IV we will provide specific use cases as experiments we envision the metrics will be used in, and how to reason about their differences. Section V will discuss some limitations of our metrics and how we might improve or extend them to other domains.

II. Background

Bias and fairness in machine learning have received increasing attention in recent years. The advantages of algorithmic decision-making can be very attractive to large organizations, but there is a risk that the output of these algorithms can be unfair [8]. Unfairness can have serious perceptual and legal consequences for organizations who choose to rely on machines to make important decisions [9]. This makes it imperative that quantitative measures for bias and fairness in machine learning be defined.

Bias, discrimination, and unfairness are terms that are often used interchangeably but we would like to make a distinction to better dissect the problem. We will refer to bias as meaning the behavior of a machine learning model giving preference to one characterization over another [10]. Or put simply, having a lower error rate on one class than another. When discussing fairness in this paper, we will be referring to group fairness, which in general is concerned with outcomes for privileged and unprivileged groups [5], where a group is a protected feature of an instance from the training data characterizing the instance in some way. Typically we do not want membership in a group to affect the outcome of a prediction, e.g. considering race or gender for ranking resumes or home loan applications.

There are other accepted definitions of fairness as well [8]. Individual fairness requires that a model give similar predictions for similar individuals. Subgroup fairness, which uses notions of both individual and group fairness by holding some constraint over large collections of a subgroup. However, group fairness is the most commonly measured by metrics of fairness [8].

Both bias and unfairness can degrade the performance of a model in ways that are not well captured by accuracy, precision, and other measures of machine learning (ML) performance. Biased and unfair models can perform very well on biased or unfair data. A nuanced metric can reveal conditions under which a model's performance might be degraded by the bias of the model or its training data. Many good metrics exist for measuring the group or individual unfairness of a model, but the focus has overwhelmingly been on tasks of supervised, binary classification [9].

Substantial literature has emerged concerning algorithmic bias, discrimination, and fairness. Mehrabi et al. conducted a survey on bias and fairness in machine learning [8]. Mitchell et al. explored how model cards can be used to provide details on model performance across cultural, racial, and inter-sectional groups and to inform when their usage is not well-suited [11]. Gebru el al. proposed using datasheets as a standard process for documenting datasets [12]. Amini et al. proposed to mitigate algorithmic bias through re-sampling datasets by learning latent features of images [13]. Wang et al. designed a visual recognition benchmark for studying bias mitigation in visual recognition. [14].

Other metrics of fairness have been described in recent works [2]–[5] whose purpose is to measure unfairness in machine learning models. Measurements of fairness based on the area under the receiver operating characteristic curve (AUC-ROC) are described in [4] and expanded in [2]. These metrics measure group-wise accuracy using AUC. Prediction Sensitivity, described in [5] fills the need for a reliable measure of individual fairness, as opposed to group fairness. A common shortcoming of these metrics is that they focus predominately on binary classification [9] and are not meaningful in tasks of multi-class classification. Our proposed metrics are usable with any number of classes. Additionally, few works have studied how bias can present as unfairness and vice versa.

To the best of our knowledge, our work is among the first to propose a single metric shown to express both bias and unfairness when comparing two models. Other work has extended the definition of Demographic Parity to include multi-class classification [15] to provide a rule for optimal classification under a restraint of Demographic Parity. Our metrics are distinct from this previous work in that they pro-

vide a method for detecting issues of unfairness in class-wise error rates rather than in overall accuracy. Mutli-calibration, the distinguish-ability of multi-class predictions and ground truth to down stream observers, is being re-investigated as a measure of group fairness [16], but this is better seen as a framework for tuning fair models rather than a metric for assessing fairness.

Multi-group fairness can be as challenging as multi-class. This is particularly true when the groups potentially overlap. Recent work [17] has shown a method for producing a Bayes-optimal group-wise classifier can be generated which maximizes fairness when measured with fractional and convex metrics.

III. Proposed Metrics for Class-Wise Bias

We propose two new metrics, Combined Error Variance (CEV) and Symmetric Distance Error (SDE)s. Both measure changes in the class-wise false positive and false negative rates of two models, and each has its own advantages which will be explored in Section IV. When calculating both CEV and SDE one model is used as the base and another model as the alternative.

A. Combined Error Variance

The concept of the Combined Error Variance (CEV) metric is to measure the tendency of DNNs to sacrifice performance on one class for the benefit of others. CEV approximates the variance of the change in False Negative Rate (FNR) and change in False Positive Rate (FPR). It summarizes changes in FNR/FPR away from the model's average. Mathematically, CEV is defined as follows.

$$\delta X_{ie} = \frac{X_{ie} - \hat{X_{ie}}}{\hat{X_{ie}}} \quad (1) \qquad \delta X_{\mu e} = \frac{1}{n}\sum_{i=0}^{n}(\delta X_{ie}) \quad (2)$$

$$cev = \frac{1}{n}\sum_{i=1}^{n}(dist((\delta X_{\mu pos}, \delta X_{\mu neg}), (\delta X_{ipos}, \delta X_{ineg})))^2 \quad (3)$$

Let X_{ie} be a pair of values for the FPR and FNR for class i of the comparison model and $\hat{X_{ie}}$ be the original models FPR/FNR pair, with e indicating either the false-positive or false-negative rate. We first find the normalized change in FPR/FNR, noted as δX_{ie}, by subtracting the error rates for the two models from each other and dividing by the original. The mean change $\delta X_{\mu e}$ is found by averaging the values of δX_{ie}, keeping in mind that every δX_{ie} is the change in two values FPR and FNR. The CEV is calculated by treating each δX_{ie} as a point in a 2-dimensional space of FNR and FPR. The square of the euclidean distances between each δX_{ie} and the mean change represented by $\delta X_{\mu e}$ are summed and divided by the total number of classes n. Euclidean distance has been used as the distance measure in the results presented later, but other distance measures could still be used.

B. Symmetric Distance Error

The principle of the Symmetric Distance Error (SDE) metric is to measure another undesirable bias behavior that presents in simple models. That is, a class with more training examples or that has similar features to another class is more frequently to be chosen by the model with limited capacity. To reflect this biased behavior, SDE calculates how "far away" from balanced is the change in FPR/FNR for each single class error. Intuitively, if we make a scatter plot with changes in FPR and FNR as X and Y values, the diagonal line in that plot would be a perfectly balanced change in FPR/FNR. Therefore, the SDE can be calculated as the symmetric distance of each change to that balance line.

$$d = \frac{|a(x_0) + b(y_0)|}{\sqrt{a^2 + b^2}} \quad (4)$$

$$d = \frac{|(1)(x_0) + (-1)(y_0)|}{\sqrt{(1)^2 + (-1)^2}} = \frac{|x_0 - y_0|}{\sqrt{2}} \quad (5)$$

For a line in the Cartesian plane described by the equation $ax + by + c = 0$, the distance d from any point (x_0, y_0) can be derived from the equation in 4. In our specific context, the diagonal of the Cartesian plane (i.e. the balance line) is $x = y$ or $x - y = 0$ will represent an equal difference in FNR and FPR between two models. Given any change of FNR and FPR the symmetric distance of that change to the balance line can be calculated as:

$$sde = \frac{1}{n}\sum_{i=0}^{n}|\delta FNR_i - \delta FPR_i| \quad (6)$$

Once the symmetric distance of each change is calculated, the SDE of a model can be calculated as the mean absolute change of normalized False Positive/ False Negative rates, with the change being calculated as described in Equation 1. The $\sqrt{2}$ has been omitted from the final equation as a constant that has no effect on the meaning of the metric. This metric will therefore reveal that one model or the other is more biased toward false positives or false negatives in a class-wise fashion.

C. Normalization

Both CEV and SDE may produce a large range of values depending on the specific dataset, number of classes, and performance of the models trained on that data. While not strictly necessary, in order to make the outputs of our metrics more interpretable, we follow a procedure for normalizing their values based on a hypothetical "worst performing" model to give us a reference. To do this a set of predictions for all test instances is produced at random with all classes being equally likely, and the FPR/FNR of these random predictions is calculated. The CEV and SDE of the random predictions is generated relative to the original model. These are then used as a divisor to normalize the other CEV and SDE values of a group of models. Following this process, our metrics now indicate a change in algorithmic bias relative to a random predictor. Thus, a CEV value of 0.5 shows that the class-wise bias of model 2 relative to model 1 has increased by 50% of the change between model 1 and a random predictor.

TABLE I: Summary of datasets used in Section IV

Name	# Train Instances	Data Type	# Classes
CIFAR100	60,000	Image	100
Titanic	891	Tabular	2
CelebA	202,599	Image+Annotations	40 binary attributes

TABLE II: Low resource ImageNet models from TIMM github [18]. Top1/Top5, input image sizes, and parameter count listed. Index corresponds to axis labels in Figure 2

index	model	top1	top5	img size	params x10^6
0	efficientnet_b2	80.608	95.310	288	9.11
1	efficientnet_b1	78.792	94.342	256	7.79
2	efficientnet_b1_pruned	78.242	93.832	240	6.33
3	mobilenetv3_large_100_miil	77.914	92.914	224	5.48
4	mobilenetv2_120d	77.294	93.502	224	5.83
5	mobilenetv3_large_100	75.768	92.540	224	5.48
6	mobilenetv3_rw	75.628	92.708	224	5.48
7	mobilenetv2_110d	75.052	92.180	224	4.52
8	pit_ti_distilled_224	74.536	92.096	224	5.10
9	deit_tiny_distilled_patch16_224	74.504	91.890	224	5.91
10	mobilenetv2_100	72.978	91.016	224	3.50
11	resnet18	69.758	89.078	224	11.69

IV. Example Applications

We have explored several applications of CEV and SDE for comparing the performance of two models. While we don't believe this list is exhaustive, in this section we illustrate several scenarios where our proposed metrics can be used. The datasets used in these demonstrations are summarized in Table I. We group these applications into two categories:

1) We demonstrate using CEV/SDE in the context of informing and selecting from any number of trained low resource models to replace a higher capacity model.
2) Evaluating group fairness. We demonstrate how CEV/SDE can be used to measure relative bias w.r.t protected groups. We also compare our results to existing binary classification fairness metrics and demonstrate the use of our metrics on multi-class data.

A. Model Selection

Bias is an important consideration when selecting a pre-trained model from one of the dozens which are available in many problem spaces. Here we see how one might use CEV/SDE to detect and avoid a model more biased than models of similar accuracy. For this example, we have selected a set of models from the TIMM model repository [18] that have between 3.5×10^6 and 11.7×10^6 parameters. Each model has been pre-trained on the Imagenet dataset [19]. Table II lists the specific models, their top1/top5 accuracy, image input size, and number of parameters. We have constructed heat maps of the CEV/SDE values by calculating the interaction between each model and building an adjacency matrix. In both Table II and Figure 2 we sort the models by Top-1 accuracy. With our constructed matrices we can quickly observe that mobilenetv3_large100 on row 5 column 5 stands out clearly in the CEV/SDE matrices. We see that although the model has comparable accuracy and parameters to mobilenetv3_rw and mobilenetv2_110d, it is actually measured to have worse trade-offs of FPR/FNR w.r.t to the tables best model in terms of accuracy efficientnet_b2, and is no better or worse than several of the next several models on our accuracy sorted list. CEV and SDE have prevented us from making a poor selection with relative ease. We find accuracy alone is a poor indicator of model quality. For example mobilenetv3_large_100 (#5) has significantly different per-class accuracy compared to models with similar top-1 accuracy.

TABLE III: Comparison of Error Rate Equality Difference(ERED) [4], Difference in Expected Value(DEV) [3], and proposed CEV/SDE on the Titanic dataset [20]. CEV/SDE are calculated w.r.t the whole dataset errors, and given protected group. Values are averaged over 5 runs of train/test.

Model	Our Metrics				Existing Metrics			
-	CEV		SDE		ERED		DEV	
-	All→Men	All→Women	All→Men	All→Women	FPED	FNED	DIMS	DIAMR
NN	0.013557	0.012737	0.115002	0.093218	0.548443	0.458016	-0.269742	0.288790
SVM	0.012089	0.000736	0.109744	0.027081	0.412500	0.593508	-0.067460	0.491071
GTB	0.000107	0.000941	0.010341	0.030619	0.458462	0.513932	-0.193700	0.364831

B. Fairness

Fairness is often defined as the ability of a model to classify all groups within the testing data equally well. For example, a model trained to recognize human faces should be equally good at recognizing the faces regardless of demographic traits (e.g race, gender, age). Unfortunately, unintentionally biased data collected in real-world datasets and even train methodologies can cause undesired performance in models. Our metrics were developed specifically to measure the bias of classifiers, but we will demonstrate they may also be used for measuring fairness as well. Importantly, this methodology allows the metrics to measure fairness in multi-class examples.

To measure bias with CEV or SDE, one model is compared to another. This process can be adapted to measure fairness by comparing a model's performance on its test data to its performance on a subset of its test instances. For this purpose, we select from specific protected attributes and calculate bias with respect to the groups. A large value in CEV or SDE indicates that per-class bias is increased for one group in the data, and that the model's performance is lower for that group.

1) Binary Classification: To demonstrate measuring fairness in binary classification, we trained several common machine learning models on the Titanic dataset [20]: a shallow neural network (NN), a support vector machined (SVM), and a gradient tree boosting classifier (GTB). This dataset offers information about Titanic passengers with the labels Survived and Did Not Survive. The sex of each passenger is included as a feature of each instance. Sex was excluded in the model training and used later for group-wise fairness testing. These metrics are presented along with the False Positive Equality Difference(FPED), False Negative Equality Difference(FNED) [4], Difference in Mean Scores(DIMS), and Difference in Average Model Residuals(DIAMR) [3] in Table III.

The four metrics presented for comparison are all zero for perfectly fair predictions. The relatively small value generated for each of the eight metrics is an effect of the small size of the dataset. The fact the FPED, FNED, DIMS, and DIAMR

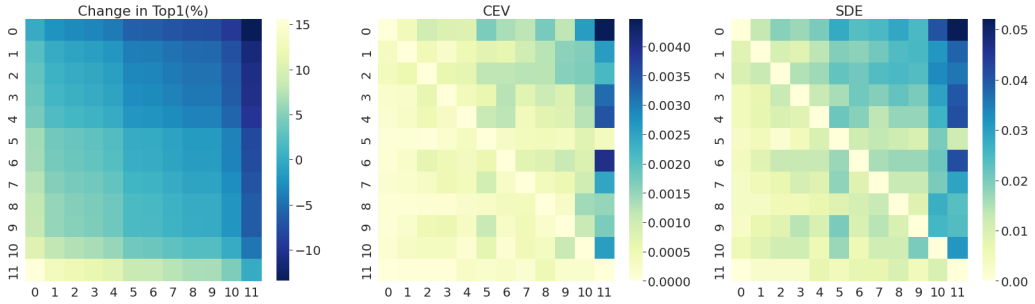


Fig. 2: Change-in-Top1, normalized CEV, and normalized SDE adjacency matrices of models listed in Table II. Each entry displays the given metric calculated for the columns model w.r.t the rows model. Reading the table row-wise you see trade offs going from the row model to another. Reading the column you see the trade offs for other models going to the column model. Higher values CEV/SDE (shown by darker cells) indicate moving towards a more biased model.

are not 0, shows that some unfairness has been learned by our neural network. The differences in the CEV and SDE scores moving from all data to men only, and all data to women only also indicates biased and unfair performance by the classifier. So we can confirm that in tasks of binary classification, our new metrics conform to the established work in the field of fairness. But as will be shown in Section IV-B2, CEV and SDE are not limited to the analysis of binary labels.

2) Multi-Class Classification: We have claimed that CEV and SDE can be used to measure fairness in multi-class classification. We will now demonstrate that process using the CelebA dataset [21]. This data contains several thousand images of celebrates and public figures with 40 binary attributes. We have selected a subset of attributes representing hair color to serve as training labels. We then trained a ResNet34 image recognition model to identify the hair color of the image subjects. From the remaining provided attributes, we have identified several to serve as protected groups ("Attractive", "Male", "Pale Skin", "Young"). As these labels come from what might be described as "privileged", we also consider subsets formed from the conjugate of these labels. It is worth noting that the conjugate does not imply the opposite. For example, the absence of a Pale Skin label does not explicitly mean dark skin but would contain all of those examples.

The results are contained in Figure 3 and Table IV. We find that groups "Male" and "Pale Skin" have the highest Top-1 accuracy. However, we also find they have high levels of class unfairness. Specifically for Male, our model is far less likely to correctly identify Male as having Blond Hair, and more likely to incorrectly guess they have Gray Hair. Meanwhile, "Not Pale Skin" has lower accuracy, but the accuracy and FPR/FNRs are much closer to the average of the model as a whole. This is easily visible in Figure 3. This unevenness is neatly captured by the corresponding CEV and SDE values or the groups in our data.

TABLE IV: Protected Attributes performance metrics for ResNet model trained on CelebA dataset.

Protected Attribute	Top-1	CEV	SDE	Change in FPR	Change in FNR
Full Test Set	0.9212				
Attractive	0.9222	0.0015	0.0331	-31.0809	80.4380
Male	0.9225	0.1413	0.2205	12.8440	77.8003
Pale Skin	0.9224	0.0035	0.0465	-43.8572	-33.9335
Young	0.9215	0.0002	0.0082	-27.6765	150.5065
Not Attractive	0.9208	0.0034	0.0493	45.6423	6.8297
Not Male	0.9207	0.0053	0.0562	1.2762	47.2981
Not Pale_Skin	0.9207	0.0000	0.0021	1.9565	1.4648
Not Young	0.9213	0.0035	0.0313	146.3057	0.2381

V. Discussion and Limitations

As with any metric, it is also important to remember that CEV and SDE are only meaningful in context. A higher value for CEV indicates that the second model has a higher class-wise bias. A higher value for SDE indicates that the second model is skewing towards false positives or false negatives. Either behavior represents a degraded real-world performance for a model in a way that may not be captured by accuracy or precision, as demonstrated in Section IV.

Data that meaningfully describes the real world is often multi-class. While it is true to that multi-class classification can be re-framed as many binary classification problems, re-framing a problem as 100 or 1,000 one-vs-each problems would only serve to make reasoning about the implications much more difficult. We believe CEV and SDE are applicable to many real-world problems completely ignored by their binary cousins.

CEV and SDE can be used to measure the fairness of a machine learning model, but only group fairness. Individual fairness, which is defined as the degree to which similar individuals are classified similarly, is not measured in any of the use cases presented in Section IV.

We have not found any consistent threshold that indicates by itself that a model is or is not biased. Another important note is that biased performance may be the result of algorithmic bias, or it may be a reflection of biased data and CEV/SDE alone cannot determine its source. Despite these limitations, CEV and SDE reliably indicate that one model is more or less biased than another. As concluded in [3], "Fairness metrics in machine learning must be interpreted with a healthy dose of human judgment."

CEV and SDE are calculated w.r.t to some other classifier and only classifiers. As such they are not suitable for every

Fig. 3: Change in FP/FN rate for protected subgroups of ResNet model trained on CelebA dataset. Change is calculated w.r.t the to complete validation set

situation. However, we believe they provide a good starting point for the community to begin to address measuring more sophisticated machine learning tasks. Additionally, we endeavor to extend the concepts of CEV/SDE to other tasks (e.g. image segmentation), which are harder still to quantify. We also believe our insights from CEV/SDE can be used to create stand-alone metrics to measure bias and fairness without making direct model comparisons.

VI. Conclusion

Unfairness is a persistent and difficult problem in machine learning. Bias is more quantifiable but just as dangerous to the reliable performance of machine learning models in the real world. In this paper, we have introduced two new metrics: CEV and SDE. These metrics can reliably reveal that a model is more or less biased compared to another model. We have also demonstrated that these new metrics can be used to measure the fairness of a model used for classification. Importantly, these metrics are meaningful when used with multi-class data, even with a very large number of classes.

References

[1] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, G. B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, *et al.*, "Quantitative analysis of culture using millions of digitized books," *science*, vol. 331, no. 6014, pp. 176–182, 2011.

[2] D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman, "Nuanced metrics for measuring unintended bias with real data for text classification," in *Companion proceedings of the 2019 world wide web conference*, pp. 491–500, 2019.

[3] J. H. Hinnefeld, P. Cooman, N. Mammo, and R. Deese, "Evaluating fairness metrics in the presence of dataset bias," *arXiv preprint arXiv:1809.09245*, 2018.

[4] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman, "Measuring and mitigating unintended bias in text classification," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 67–73, 2018.

[5] K. Maughan and J. P. Near, "Towards a measure of individual fairness for deep learning," *arXiv preprint arXiv:2009.13650*, 2020.

[6] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[7] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[8] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *arXiv preprint arXiv:1908.09635*, 2019.

[9] S. Caton and C. Haas, "Fairness in machine learning: A survey," *arXiv preprint arXiv:2010.04053*, 2020.

[10] R. J. Mooney, "Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning," *arXiv preprint cmp-lg/9612001*, 1996.

[11] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229, 2019.

[12] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for datasets," *arXiv preprint arXiv:1803.09010*, 2018.

[13] A. Amini, A. P. Soleimany, W. Schwarting, S. N. Bhatia, and D. Rus, "Uncovering and mitigating algorithmic bias through learned latent structure," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 289–295, 2019.

[14] Z. Wang, K. Qinami, I. C. Karakozis, K. Genova, P. Nair, K. Hata, and O. Russakovsky, "Towards fairness in visual recognition: Effective strategies for bias mitigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8919–8928, 2020.

[15] C. Denis, R. Elie, M. Hebiri, and F. Hu, "Fairness guarantee in multi-class classification," *arXiv preprint arXiv:2109.13642*, 2021.

[16] S. Zhao, M. Kim, R. Sahoo, T. Ma, and S. Ermon, "Calibrating predictions to decisions: A novel approach to multi-class calibration," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[17] F. Yang, M. Cisse, and S. Koyejo, "Fairness with overlapping groups," *arXiv preprint arXiv:2006.13485*, 2020.

[18] R. Wightman, "Pytorch image models)." https://github.com/rwightman/pytorch-image-models, August 2021.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[20] T. C. Frank E. Harrell Jr., "Titanic dataset," oct 2017.

[21] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.