2024 IEEE 25th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM) | 979-8-3503-9466-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/WoWMoM60985.2024.00028

Multiclass AUC for Comparison of Identification Effectiveness Across Classification Set Sizes

Mark Roman Miller
Department of Computer Science
Illinois Institute of Technology
Chicago, USA
mmiller30@iit.edu

***Abstract*—Virtual and augmented reality devices track the body motion of users because it is fundamental to rendering virtual content anchored to space. However, this same body motion data can be used as a biometric to identify users. Research on the effectiveness of these biometrics are often difficult to compare between because the common evaluation metric, rank-1 accuracy, is relative to the number of identities within the set. In this work, I motivate, select, and justify the use of a previously introduced classification model evaluation metric, *multiclass AUC,* that is invariant to the number of classes (i.e., individuals) being identified, producing more effective comparisons across disparate datasets, activities, and participant pool sizes. I also generalize this metric with regular rank-1 accuracy to produce N-class accuracy, allowing future work to compare to past work when multiclass AUC is not reported. The common use of this metric will allow a finer view of patterns in identifiability of this motion data, ultimately resulting in clearer research conclusions when comparing across works.**

***Index Terms*—virtual reality, motion as biometric, multiclass AUC,**

I. Introduction

Virtual and augmented reality devices track the body motion of users because it is fundamental to rendering virtual content anchored to space. However, this same body motion data can be used as a biometric to identify users [1]–[4]. Identification-focused works almost exclusively use accuracy for the model's evaluation metric. The benefits of accuracy as a metric include its ease of interpretation and its directness to the question at hand - a model with less accuracy is obviously less identifiable, and vice versa.

The focus of these previous works on *rank-1 accuracy* - the proportion of times the model is able to predict the identity from a sample given exactly one guess - implies that the accuracy is sensitive to the number of classes the model must distinguish between. Multiple works provide evidence of this effect of number of classes on rank-1 accuracy, even if the same data distribution and identification techniques are used [1], [3], [5]. This effect is intuitive: if there are more options, there are more ways for the model to be incorrect and the same number of ways for the model to be correct. This effect of the number of classes on accuracy can make synthesis of findings across works difficult, as the classification can vary as much as four orders of magnitude (e.g., 5 in [5] to over 50,000 in [6]).

To address this concern, I specify a criteria for an evaluation metric that is invariant to class size, identify a metric known to the literature that fits this criteria, justify this metric with respect to the criteria, and then introduce a generalization of this metric to allow comparisons from new work to previous work. With the effectiveness of motion as a biometric having been convincingly established [6], the future work in this domain will be exploring the boundaries and conditions in which motion can be used as a biometric. This work enables these comparisons to be made more effectively.

II. Multiclass AUC

The task this work is concerned with is determining an appropriate evaluation metric that is robust to varying numbers of classes in a multiclass classification problem. The criteria is that the desired metric should produce the same value regardless if it is computed upon the full set of classes or computed as the average of randomly chosen subsets of classes of any size. More formally, let $\mathcal{C}$ represent a classification problem whose elements $C \in \mathcal{C}$ are sets containing individual members of the class C. Define an ideal evaluation metric $\mathcal{M}$ such that the evaluation $\mathcal{M}(f, \mathcal{C})$ computed from the prediction function f and the classification problem $\mathcal{C}$ is equal to the expected value of the evaluation $\mathcal{M}(f, \mathcal{C}')$ for a randomly chosen combination of classes $\mathcal{C}'$ of a given size N, uniformly randomly selected from the classes in $\mathcal{C}$. Numerically, this is:

$$\mathcal{M}(f, \mathcal{C}) = \binom{|\mathcal{C}|}{N}^{-1} \sum_{\mathcal{C}' \subseteq \mathcal{C}, |\mathcal{C}'|=N} \mathcal{M}(f, \mathcal{C}')$$

A metric previously known to the literature, *multiclass AUC*, defined by Hand and Till [7], fits this criteria. Multiclass AUC can be described as the average of the pairwise separability between classes. In the original work, Hand and Till extend area-under-the-curve (AUC), the well-known measure of separability, to the multiclass case. AUC can be expressed as the probability that a randomly selected member a of class A will be larger than a randomly selected member b of class B according to the value of the binary prediction function f_{binary} meant to separate the two. This can be easily computed in closed form as

$$AUC = \frac{1}{|A||B|} \sum_{a \in A,\ b \in B} \mathbf{1}[f_{binary}(a) > f_{binary}(b)]$$

979-8-3503-9466-5/24/$31.00 ©2024 IEEE
DOI 10.1109/WoWMoM60985.2024.00028

where $\mathbf{1}$ is the indicator function. Multiclass AUC extends this definition provided a multiclass prediction function f that specifies values $f(m, C)$ for each combination of member m and class C. From this, Hand and Till [7] define the multiclass AUC for a given prediction function f and set of classes $\mathcal{C}$ to be the average of separabilities of one class from another for all pairs of classes in the model:

$$\frac{1}{|\mathcal{C}|(|\mathcal{C}|-1)} \sum_{A,B \in \mathcal{C},\ A \neq B} \frac{1}{|A||B|} \sum_{a \in A,\ b \in B} f(a, A) > f(b, A)$$

As a sketch for the proof that metric fills the criteria above, consider that this metric produces a separability value for each ordered pair of classes, independent of the other classes present. Due to the symmetry of classes in being selected within the final set, each class and class pair is weighted equivalently in the averaging process. By linearity, the average of the final values for a given class size can be understood as the average of all pairwise values, which produces the same value as evaluating the full set of classes.

Hand and Till note that this metric weights the separability of each pair of classes equally regardless of the number of samples in the classes, which may not be appropriate if certain classes (users) are determined *a priori* to be more likely than others. However, in the context of estimating the effectiveness of a biometric, priors are not traditionally included. This can be explicitly done to ensure equality between users, i.e., equally weighing errors regardless of who is being misidentified, or it can be implicitly done by including the same number of training instances per class (user) [2], [3], [6]. Additionally, this is not an estimate of the accuracy attained by the same training process upon a smaller data set constructed in the same class-reduction process, but is instead an estimate based upon the model after training.

III. Accuracy limited to an N-class testing set

While multiclass AUC is a good multiclass evaluation metric for future work, there are no works in this space that currently use it. In order to allow comparisons to be drawn from this work to previous work, I define N-class accuracy as a generalization of both rank-1 accuracy and multiclass AUC.

This metric may be narrated as a prediction task in which there is a model and a set of N potential classifications, a subset of all the classifications the model could make. First, the model proposes its classification, and if the classification is outside this subset, the model is asked to provide its next best classification. This process only ends when the model gives a predicted classification within the set of potential classifications.

To derive this formula, consider the probability

$$P[\arg\max_{C \in \mathcal{C}'} f(a, C) = A]$$

for a randomly selected subclassification $\mathcal{C}' \subseteq \mathcal{C}$, $|\mathcal{C}'| = N \leq |\mathcal{C}|$, that sample a known to be from class A is predicted correctly. For a sample to be correctly classified, the random selection of classes within this set of N potential classifications must avoid all classes that would trip up the prediction for a given sample a whose true class is A. The number of these 'error classes' is

$$N_{error} = \sum_{C \in \mathcal{C}} \mathbf{1}[f(a, C) > f(a, A)]$$

. The general expression for a sample a to be correctly classified in an N-class testing set is a simple combinatorics expression:

$$P[\arg\max_{C \in \mathcal{C}'} f(a, C) = A] = \frac{\binom{|\mathcal{C}| - N_{error}}{N}}{\binom{|\mathcal{C}|}{N}}$$

Then, by linearity of expectation, the accuracy for the whole model across all selections of $\mathcal{C}' \subset \mathcal{C}$ is equal to the mean of each sample's accuracy, and so the end result is simply the mean of the expression above across all sessions.

When $N = 2$, N-class accuracy is the special case of multiclass AUC, and when $N = |\mathcal{C}|$, N-class accuracy is the special class of standard rank-1 accuracy.

IV. Conclusion

In this work, I recommend and justify the use of multiclass AUC as a preferred metric when comparing accuracies across different datasets. I also propose N-class accuracy, a generalization between multiclass AUC and standard rank-1 accuracy. The goal is to make trends in accuracy clearer with a simple read through the literature. To continue to explore the benefits and risks of VR-captured motion as a biometric, we need to understand the boundaries and conditions of identifiability, which in turn requires and insightful comparisons based on previous literature.

References

[1] K. Pfeuffer, M. J. Geiger, S. Prange, L. Mecke, D. Buschek, and F. Alt, "Behavioural Biometrics in VR: Identifying People from Body Motion and Relations in Virtual Reality," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: ACM, 2019, pp. 110:1—-110:12. [Online]. Available: http://doi.acm.org/10.1145/3290605.3300340

[2] A. G. Moore, R. P. McMahan, H. Dong, and N. Ruozzi, "Personal identifiability and obfuscation of user tracking data from VR training sessions," *Proceedings - 2021 IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2021*, pp. 221–228, 2021.

[3] M. R. Miller, F. Herrera, H. Jun, J. A. Landay, and J. N. Bailenson, "Personal identifiability of user tracking data during observation of 360-degree VR video," *Scientific Reports*, vol. 10, no. 1, pp. 17 404–17 413, 2020. [Online]. Available: https://doi.org/10.1038/s41598-020-74486-y

[4] V. Nair, G. M. Garrido, D. Song, and J. F. O'Brien, "Exploring the privacy risks of adversarial VR game design," *Proc. Priv. Enhancing Technol.*, vol. 2023, no. 4, pp. 238–256, 2023. [Online]. Available: https://doi.org/10.56553/popets-2023-0108

[5] X. Wang and Y. Zhang, "Nod to Auth: Fluent AR/VR Authentication with User Head-Neck Modeling," *Conference on Human Factors in Computing Systems - Proceedings*, 2021.

[6] V. Nair, W. Guo, J. Mattern, R. Wang, J. F. O'Brien, L. Rosenberg, and D. Song, "Unique Identification of 50,000+ Virtual Reality Users from Head & Hand Motion Data," Feb. 2023, arXiv:2302.08927 [cs]. [Online]. Available: http://arxiv.org/abs/2302.08927

[7] D. J. Hand and R. J. Till, "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems," *Machine Learning*, vol. 45, no. 2, pp. 171–186, 2001.