Identifying Indicators of Bias in Data Analysis Using Proportionality and Separability Metrics

Madeleine Schneider[1], Jonathan Grinsell[1], Travis Russell[1,2], Randall Hickman[1,2], and Robert Thomson[1,2]

[1] United States Military Academy, West Point NY 10996, USA
[2] Army Cyber Institute, West Point NY 10996, USA

Abstract. There have been a number of high-profile cases of artificial intelligence (AI) systems making culturally-inappropriate predictions, mainly due to imbalanced training data and the overall black-box nature of modern deep learning algorithms. In this paper, we consider metrics for analyzing the data to determine the characteristics which may result in a biased model. Specifically, we look at a combination of separability between and proportionality of clusters within a given dataset to identify the presence of implicit biases. We measured the effectiveness of Alpha Diversity, entropy, and euclidean norm as a measure of proportionality. Manually permuting the MNIST and fashion MNIST dataset, we found all three scores strongly correlated with model accuracy. Silhouette scores were used as the separability metric and showed limited but noticeable correlation with predicted accuracy.

Keywords: Machine Learning · Bias · Separability · Proportionality

1 Introduction

Machine Learning is used in countless applications today. One common type of machine learning uses a method called supervised learning. With supervised learning, AI must be taught different characteristics of categories through pre-labeled data. In this learning phase, the AI is fed a set of inputs with information about the input's features. It is also fed a set of corresponding categories. Thus, during the training phase, the algorithm knows what category an input belongs to. This allows it to learn how different features categorize a given input. Once it has mapped out the differences in categories, based on the probability of certain features and feature combinations, it can begin categorizing unlabeled data. One issue with supervised machine learning is that it can only be as good as its training data. Biased training data, whether that be in content or form, can result in disastrous outcomes for machine learning and has already collectively costs businesses millions of dollars [9].

Biased training data can be particularly problematic when machine learning is being used to make important decisions such as diagnosing medical problems, identifying criminals, or determining the amount of credit someone is allowed. Currently, most problems with bias in machine learning models are not found

until after the model has been trained, or unfortunately, after the model has been put into use via a product that affects everyday lives. Depending on data size, training a model can take immense computational power. Waiting to search for bias until after the learning phase will cost time and money. Past research has worked to uncover examples of bias such as in facial recognition for different ethnicities[3] or sought to determine methods for discovering just class imbalance in data [4]. Additional work has considered finding bias in a final algorithm [6] or via unsupervised clustering of data [10].

This paper expands on these ideas to find a way to screen data before training and to consider additional problems beyond class imbalance that can help to spot potential areas of bias before an algorithm is trained. It will specifically look at two components of biased data. These components are proportionality and separability.

2 Bias Detection Methods

In order to consider proportionality and separability and the effects they have on the success of a machine learning algorithm, there must be clear ways to measure these components and the outcome.

2.1 Separability

Machine learning algorithms rely on separability to effectively classify. We consider two qualities related to separability: inter-class separability and intra-class cohesion. Machine learning algorithms work best when different classes are highly separable, meaning the characteristics of the classes are very different. They also work best when the data points belonging to one class are very similar, or cohesive. Inter-class separability looks at how easily different classes can be separated and intra-class cohesion considers how close data in the same class is to each other.

One method of measuring this combination of separability between classes and unity within a class is by using a silhouette score[2]. Silhouette scores have been particularly useful in unsupervised learning when trying to determine the appropriate number of clusters. The score gives an indicator of how close points in one cluster are to other clusters, which for unsupervised learning may indicate that a certain point was misclassified [1]. This metric is particularly useful for considering inter-class separability and intra-class cohesion.

Silhouette scoring is useful in that it considers all data points, but it is somewhat computationally intensive, running in quadratic time. Thus, it may be useful to instead consider a similar approach to a simplified silhouette method.

With the simplified silhouette method, inter and intra-cluster scores are based on the k-mean center of an unsupervised k-means cluster instead of the average distance to all other points in a cluster [11]. Because the data being considered for this paper is labeled, considerations can be made by simply finding the exact centroid (or mean) of each class. This paper further simplifies the metric by

creating a separability heuristic that considers just inter-class distance by looking at the distance between the centroids of each cluster.

2.2 Proportionality

Machine learning algorithms also have a hard time effectively classifying if there is class imbalance. It is widely accepted that algorithms will perform most effectively if classes have an even distribution in the learning data. A useful measurement of evenness of proportionality comes from the study of habitat richness and evenness in species. Alpha diversity (1) gives a diversity score that specifically "penalizes" if the proportion of one species is far away from the even proportion of $\frac{1}{C}$, where C is the number of species [4]. Applying this to machine learning, where C is the number of classes and p_i is the proportion of class i in the dataset, an alpha diversity can be calculated as follows:

$$\alpha = \sum_{i=1}^{C} p_i^2 \tag{1}$$

With this formula, a perfectly evenly split dataset would have an alpha diversity score of $1/C$. The worst score possible would occur if all the data came from one class and would give a score of 1.

Another possible metric to consider separability is entropy (2). Entropy is another measure of diversity. Similarly to alpha diversity, this metric is used as a diversity index in ecology and measures not only the amount of species, but how evenly the species are split [8]. Entropy, however, is also already an important metric in machine learning where it is used as an impurity measure [7].

$$e = -\sum_{i=1}^{C} p_i log(p_i) \tag{2}$$

With this formula, a more evenly split data set will have a higher entropy and a dataset of 1 class has a score of 0.

3 Tests

3.1 Data

Testing considered the MNIST dataset and fashion MNIST dataset. MNIST and Fashion MNIST are widely used machine learning datasets that each hold 60,000 images. MNIST contains images of handwritten digits and has been been used extensively in machine learning research on optical character recognition [5]. Fashion MNIST contains images of clothing items and was made as a direct "drop in" replacement for the MNIST dataset. It was important to use Fashion MNIST to show that the results of the two metrics were not only applicable to MNIST, but have potential to give information about many different datasets.

3.2 Method

For testing, every class combination pair of two classes was considered, although many more class combinations exist within the dataset. The separability score of each pair was computed and models were trained and tested for the proportions 50/50, 45/55, 40/60, 35/65, 30/70, and 25/75.

An "off-the-shelf" model given by Tensorflow was used. The model had 1 Flatten layer which took the two dimensional pixel array and flattened it into a one dimensional vector. Following this there was one Dense layer of size 512. After that there was a Dropout layer to prevent over-fitting followed by another Dense layer of size 10, the number of classes. There was one modification from the basic MNIST model. An early stopping metric was used which stopped training if the loss value stopped improving between two epochs. This also prevented the model from over-fitting to the training data.

Accuracy testing was done on balanced proportions of the two number classes to see how well the machine learning performed with different separability and proportionality scores.

3.3 Results

MNIST Separability and accuracy appear somewhat correlated. For every proportion combination, there was a correlation coefficient of just above .5. This indicates that higher separability tended to result in better accuracy.

The weak correlation may indicate that the current metric to evaluate separability is not precise enough to adequately capture the information available from understanding inter-class separability. Also, given that the current metric does not consider intra-class separability, it is expected that more information is needed to create a highly correlated metric.

The connection between proportionality and accuracy appears to be more correlated. The alpha value has a correlation of -0.986 and entropy has a correlation of 0.986.

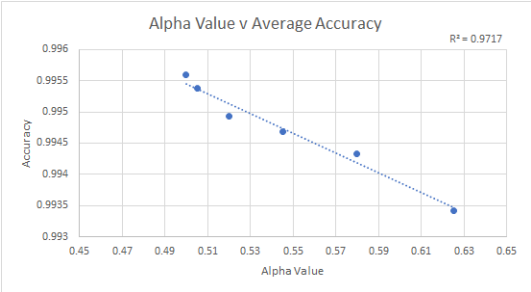

Fig. 1. MNIST Alpha Value compared against Average Accuracy. As seen, there is a negative correlation -.986 implying that the further an alpha value is from a perfectly balanced dataset, the lower the expected accuracy.

Fashion MNIST For Fashion MNIST, separability and average accuracy are more correlated than they were for MNIST. For every proportion combination, there was a correlation coefficient of just below .8. Thus, higher separability tended to result in better accuracy.

Figure 4 shows a clearer linear trend for Fashion MNIST than with MNIST, however a linear model still does not fully describe the interaction between separability and accuracy. This figure suggests that at a certain separability, a further increase does not greatly affect expected accuracy. Similarly to MNIST,

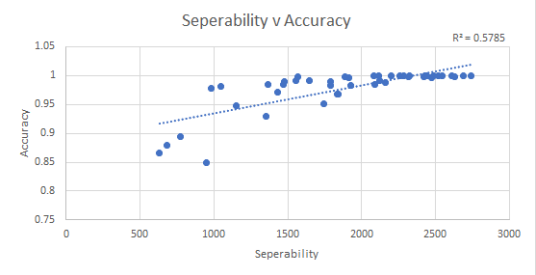

Fig. 2. Fashion MNIST Separability of Evenly Split Data v Accuracy. There is a positive correlation .751 implying that the higher the separability, the higher the expected accuracy.

for Fashion MNIST proportionality had a much higher correlation with average accuracy than separability had. The correlation coefficient for alpha value is -0.981. Entropy has a correlation of 0.982.

4 Discussion

For both MNIST and Fashion MNIST, there is an observable, albeit not high correlation between separability and average accuracy. Given that the metric was highly simplified and only considered inter-cluster separability, it is expected that a more complex metric may result in more powerful correlations. All metrics for proportionality, however showed a strong correlation. As expected, a more balanced dataset, on average, resulted in higher accuracy.

When considering the metrics together, there is a weak, but still noticeable trend. Higher separability with more balanced data generally gives higher accuracy. These metrics show how certain characteristics of data, such as separability and proportionality, can make an impact on how well a model is able to determine different classes. This helps make the outcome of a model more explainable and can help prevent programmers from using data that is deceivingly biased.

With the current metrics, all computations run in linear time. Additionally, finding the proportions of each class, to run the proportionality metrics, and finding the center point of each class can be done simultaneously. After finding the proportions and the center point, calculating all of the metrics is solely a

function of the number of classes. In total, for p data points and c classes, finding both separability and proportionality is run in roughly $O(p + 2c)$ or linear time. Thus, scalability should not be a tremendous issue. Because machine learning itself runs through each data point multiple times, if a programmer has the computational power to create an algorithm, they surely have the computation to run the metrics. It is vital that future metrics continue to run in a short enough time to ensure the metrics are scalable. By expanding to include more metrics, and by fine-tuning these metrics, it will become easier for programmers to understand their data and spot potential bias before undertaking the computationally extensive process of machine learning.

References

1. Selecting the number of clusters with silhouette analysis on kmeans clustering. scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html.
2. Bonaccorso, G.: Machine Learning Algorithms. Packt Publishing (2018)
3. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency. pp. 77–91 (2018)
4. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
5. Deng, L.: The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Processing Magazine **29**(6), 141–142 (2012)
6. Dietterich, T., Kong, E.B.: Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Tech. rep. (1995)
7. Geron, A.: Hands-On Machine Learning with Scikit-Learn TensorFlow. O'Reilly (2017)
8. Jost, L.: Entropy and diversity. Oikos **113**(2), 363–375 (2006)
9. Ovenden, J.: Bad data is ruining machine learning here's how to fix it. https://channels.theinnovationenterprise.com/articles/bad-data-is-ruining-machine-learning-here-s-how-to-fix-it
10. Thomson, R., Alhajjar, E., Irwin, J., Russell, T.: Predicting bias in machine learned classifiers using clustering. In: Annual Social Computing, Behavior Prediction, and Modeling - Behavioral Representation in Modeling Simulation Conference (2018)
11. Wang, F., Franco-Penya, H.H., Kelleher, J.D., Pugh, J., Ross, R.: An analysis of the application of simplified silhouette to the evaluation of k-means clustering validity. In: International Conference on Machine Learning and Data Mining in Pattern Recognition. pp. 291–305. Springer (2017)