What Makes a Good Diagnosis: An Algorithm to Detect Biased Training Data

Madeleine Schneider
Department of Mathematical Sciences
West Point, New York, 10996

Robert Thomson
Army Cyber Institute
Behavioral Sciences & Leadership
West Point, New York, 10996

Abstract

There have been a number of high profile cases of artificial intelligence (AI) systems making culturally-inappropriate predictions when classifying images of individuals of different races. These predictions were due in part to implicit biases within the training data. In the case of well-being, there are critical situations where AI systems can be of use, including diagnoses, treatment variation, and care decisions. The challenge of implicit bias in these critical situations is that lives are potentially on the line. Current AI approaches are generally *black-box* in that we cannot understand the features that went into a particular classification/decision. In this paper we specifically look at a combination of *silhouette score* and *alpha diversity* to identify the presence of implicit bias within a data-set. Finally, we discuss a test case where this algorithm could improve our understanding of automated diagnosis tools, specifically in diagnosing borderline personality disorder.

Introduction

There is currently ongoing research in automated machine learning techniques to supplement many complex decision-making tasks, from medical decision-making to parole decisions. A challenge in many deep learning based approaches is that their decisions are generally *black box*, that is, it is impossible to determine which factors enter into the decision. For instance, it is highly plausible that a commonly-available statistic like zip code is one of the determiners of automated parole recommendation. Socially, this would not be considered a fair determiner, but zip code does correlate with many features (e.g., economic status, overall crime rate) that confound the critical personal features of the potential parolee. In short, biased training data - whether that be in content or form - can result in disastrous outcomes.

The point here is that the algorithm is only so good as the quality of data that goes into it, and it needs to be extensively validated. Two main sources of *bias* in training data are proportionality (i.e., frequency imbalance between categories) and separability (i.e., how distinct each category is). In the case of medical diagnoses, detecting the presence of biased training is essential to avoid misdiagnoses.

Detecting the Sources of Bias

We argue that it is possible to predict the degree of bias within a given training set, and have previously [Thomson et al.2018] described a technique for predicting and mitigating which categories will exhibit the most bias.

Separability

Machine learning algorithms rely on category separability to effectively classify. [Elizondo2006] This separability comes in two forms, inter-cluster separability and intra-cluster 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. The more "different" or separable classes are from other classes and the more similar data points within a class are to each other, the easier it is for a machine learning algorithm to confidently classify [Bonaccorso2018].

One method of measuring this combination of separability between classes and unity within a class is by using a silhouette score [Bonaccorso2018]. To find a silhouette score for a data point first calculate an average intra-cluster distance. Any distance measure can be used. The following equations consider Euclidean distance. To find the average intra-cluster distance for data point $\bar{x}_i$ over all other points x_j in the class C, of size c use the equation:

$$a(\bar{x}_i) = (\sum_{j=1}^{c} d[\bar{x}_i, \bar{x}_j])/c \tag{1}$$

Next, the average inter-cluster distance from the point can be found by comparing the data point with the elements in another class, D, of size d:

$$b(\bar{x}_i) = (\sum_{j=1}^{d} d[\bar{x}_i, \bar{x}_j]))/d \tag{2}$$

then take the minimum b to consider the lowest inter-cluster distance for the data point. Finally, find the silhouette score with:

$$s(\bar{x}_i) = \frac{b(\bar{x}_i) - a(\bar{x}_i)}{max(a(\bar{x}_i), b(\bar{x}_i))} \tag{3}$$

This, gives the silhouette score of one data point in a class, and must be found across all data points in a class and averaged to get the silhouette score of the class.

Silhouette scoring is useful in that it considers all data points, but it is incredibly computationally intensive. Thus, it may be useful to instead consider a similar approach to a simplified silhouette method [Wang et al.2017]. With the simplified silhouette method, inter and intra-cluster scores are made off of the k-mean center of an unsupervised k-means cluster. 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. To find the centroid, simply add the vector values of all of the points in a class and divide by the number of points in that class.

To find the smallest inter-cluster distance for a class, find the smallest difference from the class's centroid to it's nearest neighbor centroid. With this method, a simplified silhouette score can be found for each class in linear time.

Proportionality

Machine learning algorithms also have a hard time effectively classifying if there is class imbalance. Algorithms will perform most effectively if classes have an even distribution in the learning data. A classic example of class imbalance is a machine learning algorithm that gets .98 accuracy, solely on the fact that .98 of its data is in one class. A useful measurement of the evenness of proportionality comes from describing habitat richness and evenness in animal species. *Alpha diversity* 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 [Chawla et al.2002].

Applying this to machine learning, where C is the number of classes and p_i is the proportion of class i in the data set, a "proportionality" score can be calculated:

$$p = \sum_{i=1}^{C} p_i^2 \tag{4}$$

An evenly split data-set would have an alpha diversity score of $\frac{1}{C}$. The worst score possible would occur if all the data came from one class and would give a score of 1.

Algorithm Success

By combining the simplified silhouette score and alpha diversity, it is possible to determine the most confusable classes, and thus the potential sources of bias in a given data-set. It is then possible to introspect over this more limited subset of the data to determine the optimal remediation technique (e.g., over- or under-sampling, extra training on borderline cases, or even investigating those inputs in training data itself for accuracy).

Sample Application: Mental Health Diagnoses

Mental health is a particularly troublesome application because these diseases have historically differentially affected gender and have many overlapping symptoms which makes them hard to separate. This has a large influence on treatment options. One particular example we wish to address is that of Borderline Personality Disorder. BPD presents itself with symptoms such as poor self image, impulsive tendencies, and mood swings. It is also diagnosed three times more frequently in women than men [Sansone and Sansone2011] and highly overlaps with the symptomolgy of bipolar disorder and post-traumatic stress. A consequence of this diagnosis is that BPD can be treated best with cognitive-behavioral therapies while Bipolar generally responds best to medication. Society would likely agree that we should not medicate people who can be best treated through other interventions.

One particular concern when training an algorithm to distinguish between BPD and Bipolar disorder is that there is most likely a gender-bias implicit in the training data. One possible reason for this bias is that women are more likely to seek treatment. Additionally, there are historical tendencies for male doctors to diagnose forthright women. Using machine learning for diagnosing BPD would be greatly influenced by the imbalanced gender proportionality in training data. If this data were to be fed into a machine learning algorithm for a diagnostic tool, BPD would likely be over diagnosed in women. Using alpha diversity score would readily identify these imbalances.

Additionally, classifying would be affected by separability. Using partial silhouette score would show the feature overlap from each disorder, and highlight where remediation techniques could apply to best categorize the different features of each disorder. We have previously described how clustering techniques can identify categories with overlapping features and poor performance [Thomson et al.2018].

In this poster, we will further describe some preliminary data on synthetic data-sets where we systematically vary separability and proportionality to provide an overall metric of resilience against *bias* and describe some methods for mitigating any bias that is discovered. This is just one example of how unequal representation in data and overlapping features across classes can create an unreliable machine learning model in the area of health and well-being.

References

Bonaccorso, G. 2018. *Machine Learning Algorithms*. Packt Publishing.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16:321–357.

Elizondo, D. 2006. The linear separability problem: Some testing methods. *IEEE Transactions on neural networks* 17(2):330–344.

Sansone, R., and Sansone, L. 2011. Gender patterns in borderline personality disorder. *Innovations in clinical neuroscience* 8:16–20.

Thomson, R.; Alhajjar, E.; Irwin, J.; and Russell, T. 2018. Predicting bias in machine learned classifiers using clustering. In *12th Annual SBP-BRIMS Conference*. Springer.

Wang, F.; Franco-Penya, H.-H.; Kelleher, J. D.; Pugh, J.; and Ross, R. 2017. 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*, 291–305. Springer.