loss weighting for learnability imbalance in multiclass-classification

Theodor Peifer
email: thp7219@thi.de
Technische Hochschule Ingolstadt

***Abstract*—Neural Networks have proven themselves to be powerful classification tools by solving problems in a range of domains with high accuracy. Yet this accuracy is never evenly distributed across all classes, which means that the true-positive rates of each class separately are different. This can happen even in balanced datasets since some classes are more difficult to learn by the model than others (this phenomenon is further referred to as *learnability-imbalance*). A common way to address this problem is to give a weight to the error function for each class to penalize losses of certain classes higher or lower. This research will address the determination of such weights to counteract the learnability-imbalance in balanced datasets using previously calculated evaluation scores. Therefore the goal is to find methods to lower the variance of the true positive rates of each class.**

I. Introduction

A frequent problem in classification appears when working with a dataset that has an unequal amount of samples per class. This imbalance leads the model to learn the patterns of a class with less elements worse than others. In order to prevent this, it is common to weight the error function [1] according to the size of each class, i.e. the number of samples it contains. Therefore, for every class there is a weight, which is greater the fewer elements it contains and that gets multiplied with the loss produced by its samples. Since the aim of a neural network is to minimize the overall loss and samples from a smaller class will produce a higher error, they will have a higher impact on the learning process in order to compensate for the different class sizes.

But this learnability imbalance appears also in balanced datasets for a variety of reasons, e.g. when the quality of the data of a class is lower than the rest of the data. A second reason, that will be presented later on, is that when some classes are similiar, the model can confuse their samples with each other what will often result in a lower accuracy of those classes. This issue is a light version of the imbalanced dataset problem and the inevitable product of every normal classification. Even though in many cases the learnability difference of the classes is either low or not from great interest, there can be more extreme cases where the model needs to produce fair and unbiased results. An example is *name-ethnicity classification*, where a model predicts the ethnicity of a name only by its letters [5]. Nationalities that speak the same language and therfore have similiar names (e.g. *british* and *american*) result in a lower accuracy (see figure 1) which therefore can lead to wrong or unfair interpretations when the model is used for social sciences experiments.

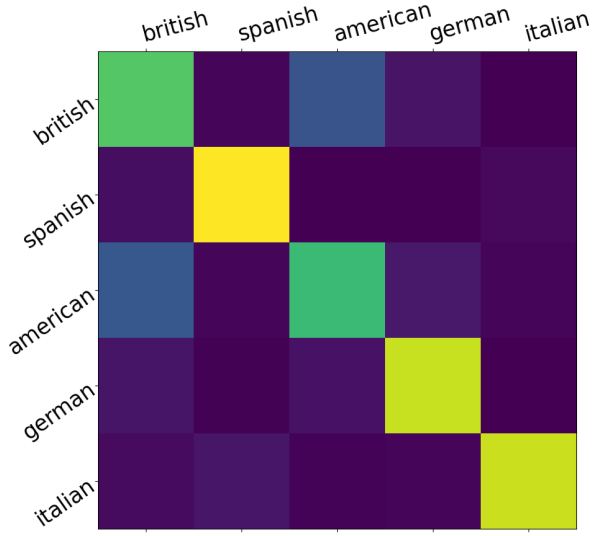

Fig. 1. confusion matrix representing the true positive distribution of the name-nationality dataset produced by a recurrent neural network

Another dataset that is a good showcase for the learnability imbalance is the CIFAR-10 [2] dataset, which consists of 32×32 RGB images of ten different classes, because it also contains similiar classes such as *dog* and *cat*. When looking at the true positive rates produced by a convolutional neural network [3] it can be seen that there are some classes that got misclassified more often (in this case bird, cat, deer and dog) which hints to a more difficult learnability. The CIFAR-10 and the name-nationality dataset will be used to run experiments, on how to prevent such variances in the evaluation scores.

II. State of the art

TODO first section here

III. Approach

TODO second section here

IV. Third section

TODO third section here

References

[1] Name Name, Name Name, Name Name (2006). Title, 24(1), 29-33.

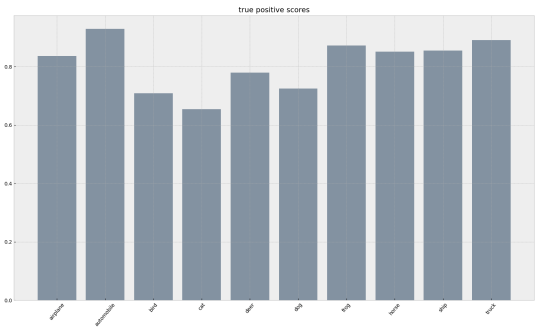

Fig. 2. true positive scores of a model trained on the CIFAR-10 dataset