CONTINUOUS WEIGHT BALANCING

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a simple method by which to choose sample weights for regression problems with highly imbalanced or skewed traits. Rather than naively discretizing regression labels to find binned weights, we take a more principled approach – we derive sample weights from the transfer function between an estimated source and specified target distributions. Our method outperforms both unweighted and discretely-weighted models on both regression and classification tasks. We also open-source our implementation of this method (LINK ANONYMIZED), providing a modular and robust software package to the scientific community.

1 MOTIVATION

Real-world datasets are heterogenous and frequently skewed. Oftentimes, the data points of interest, such as disease-positive patients in medical datasets (Rahman & Davis, 2013), are rare, and disproportionately outnumbered. Indeed, class imbalance is a well-known problem in machine learning (Japkowicz & Stephen, 2002; Japkowicz, 2000); imbalanced datasets will generally produce imbalanced models – in the case of extremely imbalanced datasets, training will result in degenerate models which opt to ignore some rare classes (Buda et al., 2018). These models can encode dataset-specific biases, and perform disproportionately better on highly represented data (Buolamwini & Gebru, 2018; Mehrabi et al., 2019). There is a rich body of literature that aims to mitigate the negative effects of dataset imbalance (Liu et al., 2008; Seiffert et al., 2009; Zhou & Liu, 2005); in practice, however, simple class balancing via reweighting is sufficient for many tasks. This method assigns sample weights inversely proportional to the amount of data in each class, thereby effectively upsampling poorly represented classes. By virtue of being simple to understand, easy to use, and effective, class reweighting is a well-worn wrench in the machine learning toolbox.

In this work, we consider a similarly simple and easy-to-use strategy to mitigate a broader type of dataset skew – imbalance of a continuous trait. A common example of this is any regression task, in which the data may be concentrated in some section of the domain. These continuous traits need not just be labels; datasets may also be biased along the axis of some continuous feature or metadata – i.e., patient age in medical datasets, or net worth in population studies. Correcting against these biases within datasets is a key step towards developing robust and unbiased models.

Our key contributions are threefold:

1. We outline a method which approximates the underlying distribution of the continuous trait, and then chooses sample weights to adjust this distribution to an arbitrary target distribution.
2. We demonstrate the performance of our method on two canonical datasets – the California housing prices dataset and the heart disease dataset – and with three classes of models – regression, random forest, and shallow neural networks.
3. We provide an open-source and modular implementation of our method.

2 KERNEL DENSITY ESTIMATES

Kernel Density Estimation (Davis et al., 2011; Parzen, 1962) is a well-known method to evaluate the probability density of a random variable given some observed samples. Formally, let $x_1, x_2, ..., x_n$

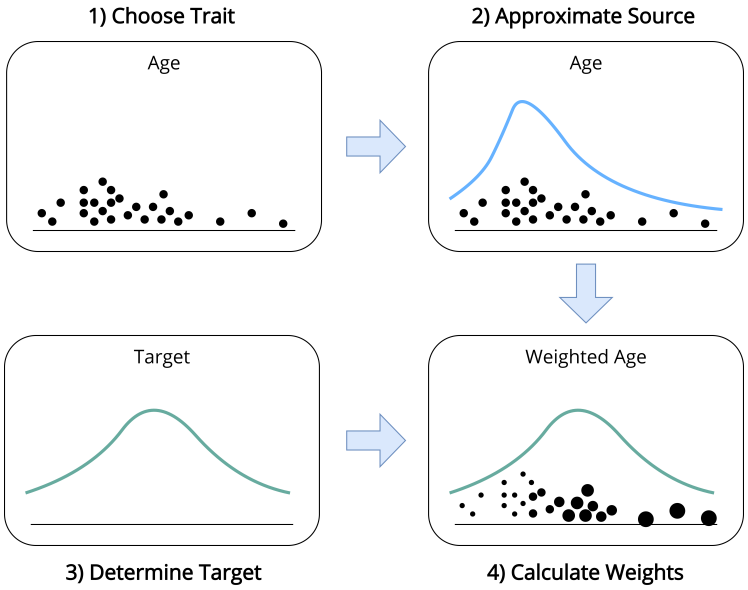


Figure 1: An illustration of continuous weighting in action.

be univariate samples drawn i.i.d. from a distribution with some density $f(x)$ at any given point x. We approximate this function f with the kernel density estimator

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where K is some non-negative kernel function, and $h > 0$ is the bandwidth parameter which smooths the resultant estimate. By convenience, the standard normal density function ϕ is commonly used as the kernel function K.

The bandwidth parameter h encodes a trade-off between the KDE's bias and variance; a common heuristic is Scott's rule (Scott, 1979), which, for a dataset of size n with dimensionality d, sets the bandwidth h as $h = n^{\frac{-1}{d+4}}$.

3 METHOD

We outline a general and flexible framework to find weights which rebalance skewed datasets containing continuous features. Our method is a simple four-step procedure, which takes a dataset, and returns weights which map this data to some target distribution.

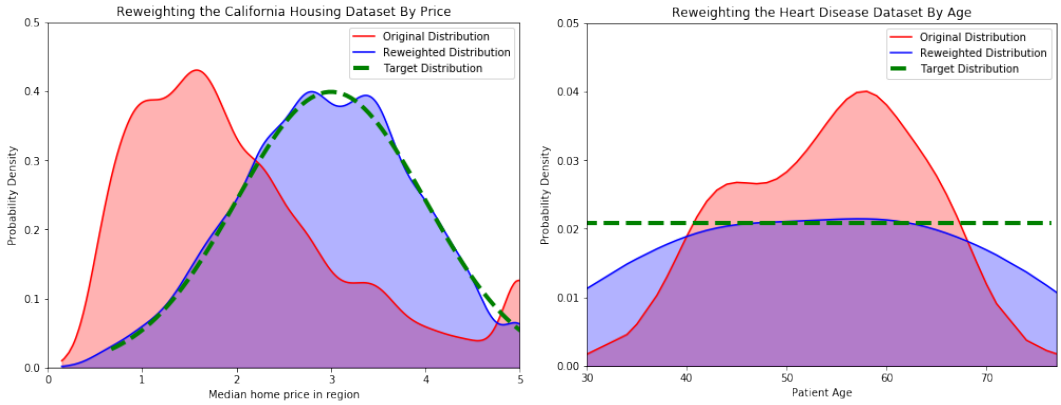
Choose a weight trait In our parlance, a weight trait is some continuous variable that captures an important feature of each data point; this is the variable we would like to weight based off of. This may be the label, a feature of the data point, or some metadata about each point. This trait may even be somewhat orthogonal to the modelling task. For example, the magnitude of each earthquake in earthquake datasets, or the volume of a given stock in financial datasets, may be traits which capture some notion of importance of a given data point, and for which we may want to weight. Choosing a weight trait may be a simple process of extracting some feature in the dataset or some corresponding data, or it may involve manual trait construction.

Approximate the source distribution We approximate the empirical distribution of weight traits with a normal kernel density estimate, with bandwidth set by Scott's rule. This produces a smoothed estimate of the underlying data distribution, which is particularly useful in case of sparsely sampled or highly skewed traits.

Determine a target distribution In this step, we determine the ideal distribution of the weight trait – the distribution of the trait in our dataset that we would like to have, for one reason or another. Generally, this target distribution can be specified by some characteristics of the problem setting or dataset. For example, if the sample is skewed from the source population, it may be prudent to reweight traits (age, income, etc.) to match the source distribution. In a different vein, the problem may contain a trait that captures some notion of importance. For example, the market size of companies may be an important trait to weight with when assembling a portfolio that focuses on a certain company size.

Determine weights Once the source and target distributions are specified, the only remaining task is to find a set of weights which transforms the source distribution into the target distribution. One simple way to do this is to set the weight on each data point to the ratio between the source density and target density evaluated at that point. Formally, for a dataset $\{x_1, \ldots, x_n\}$ with corresponding traits $\{t_1, \ldots, t_n\}$, and an approximated source probability density f_S and target probability density f_T, we calculate the corresponding weights as

$$w_i = \frac{f_T(t_i)}{f_S(t_i)}$$



(a) California housing data reweighted to $\mathcal{N}(3, 1)$ (b) The heart disease dataset reweighted to $\mathcal{U}(29, 77)$

Figure 2: Our continuous reweighting method applied to our two datasets. Larger datasets (a) are easier to reweight than smaller datasets (b).

4 Experiments

To demonstrate the practical efficacy of our method in correcting skewed datasets, we compare our weights against uniform and discretized weighting across several model classes.

4.1 Datasets

We run our experiments on two canonical datasets – one regression, and one classification.

California Housing Dataset The California Housing dataset (Pace & Barry, 1997) contains the median housing prices of Californian census block groups in the 1990 census. It consists of 20,640 data points, where each data point contains 8 numeric attributes of houses in that block, and an accompanying median home price. These attributes are the population and median income of the block, and the latitude, longitude, average occupancy, average bedroom count, average room count, and median age of homes in the block. In common usage, the natural log of the home price is used as the label. For this dataset, we simply use the target variable – the natural log of median housing prices – as the weight trait. The dataset is skewed right, and we use a unit normal target distribution centered at 3 (Figure 2a).

Model	No weights	Discrete	CWB
California Housing (R2 score)			
Random Forest regression	0.5156	0.5042	**0.5226**
Linear Regression	0.0476	0.2892	**0.3501**
Fully-connected network	0.4496	0.4237	**0.5066**
Heart Disease (AUROC)			
Random Forest classification	0.9554	0.9593	**0.9602**
Logistic Regression	0.9261	0.9223	**0.9242**
Fully-connected network	**0.9375**	0.9267	0.9324

Table 1: Experiments on both datasets using continuous weight balancing (**CWB**), discrete balancing (**discrete**) and no weights. Metrics are reported on target out-of-sample subsets; patients under 60 for the heart disease dataset, and prices above 2 for the housing dataset.

Heart Disease The Cleveland Heart Disease Database (Detrano et al., 1989), containts the clinical information of 303 patients undergoing angiography. Each patient has 14 attributes, including demographic information such as age and sex, as well as condition-specific features such as rest ECG type and maximum heart rate. The target is a binary label indicating presence of heart disease. This dataset tests the limits of our method, as it is both extremely small and highly skewed. For this dataset we use age, which is slightly skewed left, as the weight trait, and we use a uniform target distribution across the domain as the weight trait (Figure 2b).

4.2 Evaluation

We compare our continuous weights against *no weighting* and *discretized weighting*. For discretized weights, we first group *training* data into discrete bins based on a weight trait, then reweight the data in each bin until each bin has uniform representation in the training set. We evaluate the R2 score for regression and AUROC for binary classification, on an out-of-sample subset of the data. Since our goal is to enable models to do well on underrepresented subsets of our data, we evaluate model performance on underrepresented subsets of the out-of-sample data. For the California housing dataset, we examine higher housing prices (> 2), while for the heart disease dataset we examine lower age groups (< 60).

4.3 Models

We experiment with 3 classes of models: random forests, linear/logistic regression, and shallow neural networks. Our random forests use 100 estimators, and our neural networks are an ensemble of 10 feedforward neural networks, with two hidden layers containing 64 and 16 nodes, respectively. We use ReLU activations and apply dropout ($p = 0.5$) between the two layers.

5 Discussion

Table 1 outlines our method's performance. While performance gains are small in the smaller dataset (heart disease), our method significantly outperforms both discrete weights and no weighting across all models on the California Housing dataset.

In this work, we described a framework for continuous weight balancing, and assessed the performance of one simple way of doing so. While our method showed reasonable results, there are still many open questions about the best way to find continuous weights. What is the best way to approximate the source distribution? What classes of target distributions lead to the best performance, particularly with a given loss function? In extremely skewed data, how do we strike the balance between skewed models, and models which memorize specific examples, with high magnitude weights? We hope that these and other questions may be answered in future work, in order to enable the development of robust models on skewed datasets.

REFERENCES

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Richard A Davis, Keh-Shin Lii, and Dimitris N Politis. Remarks on some nonparametric estimates of a density function. In *Selected Works of Murray Rosenblatt*, pp. 95–100. Springer, 2011.

Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H. Guppy, Stella Lee, and Victor Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, 64(5):304–310, 1989. ISSN 0002-9149. doi: https://doi.org/10.1016/0002-9149(89)90524-9. URL https://www.sciencedirect.com/science/article/pii/0002914989905249.

Nathalie Japkowicz. The class imbalance problem: Significance and strategies. In *Proc. of the Int'l Conf. on Artificial Intelligence*, volume 56. Citeseer, 2000.

Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2008.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33 (3):291–297, 1997.

Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

M Mostafizur Rahman and Darryl N Davis. Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing*, 3(2):224, 2013.

David W Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979.

Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. Rusboost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40(1):185–197, 2009.

Zhi-Hua Zhou and Xu-Ying Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on knowledge and data engineering*, 18(1): 63–77, 2005.