Extreme Value Prediction in Imbalanced Regression

Yesen Chen, Zeyun Yang

1 Introduction

1.1 Motivation and importance of the problem

Data imbalance is ubiquitous and inherent in the real world. Real world data often have a long-tailed distribution that some classes have a significantly lower number of examples in the training set than other classes. Methods of dealing with imbalance are well studied for classical machine learning models. However, Research in imbalanced domain learning has almost exclusively focused on solving classification tasks for accurate prediction of cases labelled with a rare class but many real-world tasks involve continuous and even infinite target values. For example, in vision applications, one needs to infer the age of different people based on their visual appearances, where age is a continuous target and can be highly imbalanced. Treating different ages as distinct classes is unlikely to yield the best results because it does not take advantage of the similarity between people with nearby ages. Therefore, in this project, we investigate existing algorithms addressing deep imbalanced regression (DIR) on face images dataset to predict age (Yang et al.).

2 Related works

2.1 Imbalanced classification

The most straightforward and common approach is the use of sampling methods. Those methods operate on the data itself (rather than the model) to increase its balance. The sampling methods either oversample the minority or undersample the majority. For example, SMOTE algorithm generates synthetic samples for minority classes by linearly interpolating samples in the same class. The issue of class imbalance can be also tackled on the level of the classifier. In such case, the learning algorithms are modified. For example, explicitly adjusting prior class probabilities.

2.2 Imbalanced regression

Regression over imbalanced data is not as well explored. Most of the work on this topic is a direct adaptation of the SMOTE algorithm to regression scenarios. However, there exist two intrinsic drawbacks. First, it fails to take the distance between targets into account, and rather heuristically divides the dataset into rare and frequent sets, then plugs in classification-based methods. Moreover, modern data is of extremely high dimension (e.g., images and physiological signals); linear interpolation of two samples of such data does not lead to meaningful new synthetic samples.

2.3 Deep imbalanced regression

DIR aims to learn from imbalanced data with continuous targets, tackle potential missing data for certain regions, and generalize to the entire target range. In delving into deep imbalanced regression, they developed two simple, effective, and interpretable algorithms for addressing DIR: label distribution smoothing (LDS) and feature distribution smoothing (FDS), which exploit the similarity between nearby targets in both label and feature space.

2.4 Dataset

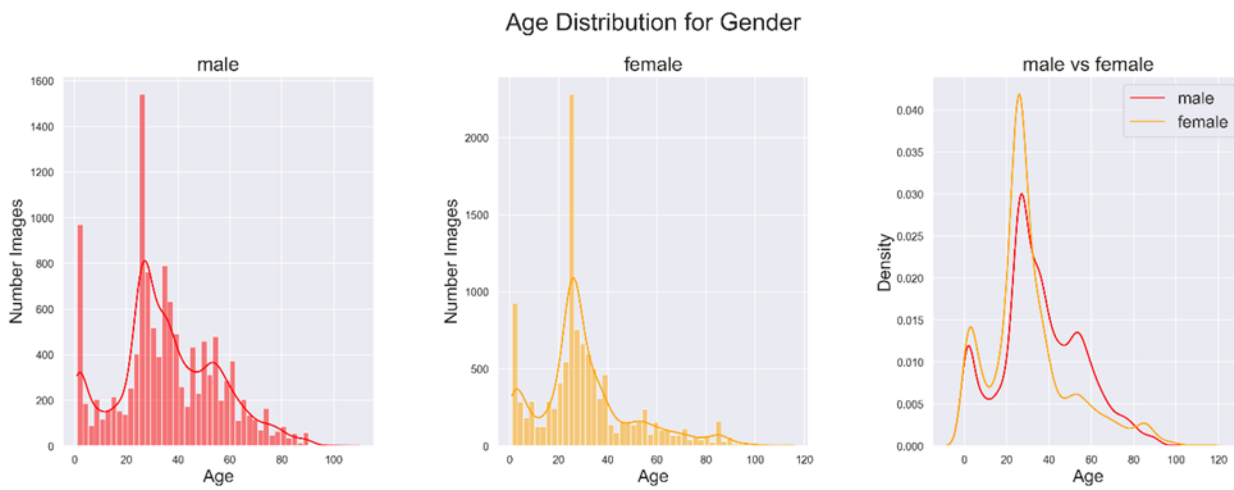
The face images dataset is called UTKFace dataset. It is a large-scale face dataset with long age span (range from 0 to 116 years old). It consists of over 20,000 face images with annotations of age, gender, and ethnicity. The images cover large variation in pose, facial expression, illumination, occlusion, resolution, etc.

3 Details of the project

3.1 Contribution of each member of the team

We worked together. The contribution of each member is half and half.

3.2 EDA



We firstly discover age distribution for gender of face images shown above. It is clear that for both genders, most people have age between 25-30. The distributions for both genders are right skewed that older people are rarer than younger people. It is consistent with the condition in the real world. For example, you hardly see people with 100 years old. Therefore, it fits the definition of imbalanced data. In addition, there are more female images than male images.

Moreover, we discover age distribution of all races shown above. It is clear that for all races, most people have age between 25-30. The distributions of all races are right skewed. The number of images of children of white race are much higher than other races

3.3 methods

After affirming that this dataset is imbalanced, we are going to explore three algorithms: simple deep regression, label distribution smoothing deep imbalanced regression and feature distribution smoothing deep imbalanced regression.

3.3.1 simple deep regression

In delving into deep imbalanced regression, they apply a 50 layers residual network as the structure of this algorithm and following two algorithms. However, due to computational limitation of our computers, we decide to choose a 20 layers residual network.

3.3.2 label distribution smoothing deep imbalanced regression

The empirical label distribution does not reflect the real label density distribution. This is because of the dependence between data samples at nearby labels. For example, images of close ages are similar. Thus, Label Distribution Smoothing (LDS) advocates the use of kernel density estimation to learn the effective imbalance in datasets that corresponds to continuous targets. Given a continuous empirical label density distribution, LDS convolves a symmetric kernel k with the empirical density distribution to extract a kernel-smoothed version that accounts for the overlap in information of data samples of nearby labels. In our project, we set the kernel to be a Gaussian kernel with size 9 and variance 1. Now that the effective label density is available, techniques for addressing class imbalance problems can be directly adapted to the DIR context. For example, a straightforward adaptation can be the cost-sensitive re-weighting method, where we reweight the loss function by multiplying it by the inverse of the LDS estimated label density for each target.

3.3.3 Feature distribution smoothing deep imbalanced regression

Continuity in the target space should create a corresponding continuity in the feature space. That is, if the model works properly and the data is balanced, one expects the feature statistics corresponding to nearby targets to be close to each other. Feature distribution smoothing (FDS) performs distribution smoothing on the feature space, basically transfers the feature statistics between nearby target bins. This procedure aims to calibrate the potentially biased estimates of feature distribution, especially for underrepresented targets. So, we have a model that maps the input data to continuous predictions. Now, FDS is performed by first estimating the statistics of each bin. Given the feature statistics, we employ again a symmetric kernel k to smooth the distribution of the feature mean and covariance over the target bins. This results in a smoothed version of the statistics. Now, with both the estimated and smoothed statistics, we then follow the standard whitening and re-coloring procedure to calibrate the feature representation for each input sample. The whole pipeline of FDS is integrated into deep networks by inserting a feature calibration layer after the final feature map. Finally, to obtain more stable and accurate estimations of the feature statistics during training, we employ a momentum update of the running statistics across each epoch.

4 Experimental results

4.1 Experiment setting

We choose Adam optimizer with momentum of 0.9 and weight decay of 0.0001 because Adam has faster computation time, and requires fewer parameters for tuning than other optimizers. We choose initial learning rate to be 0.001 and batch size to be 64. We choose epoch to be 30 because it took 2 minutes in average to train 1 epoch which is so time consuming. Moreover, the performance with epoch of 30 is already satisfied. We choose three loss functions: mean squared error (MSE) mean absolute error (MAE) and geometric mean of absolute error (G-mean).

4.2 Results and analysis

	Resnet20			Inverse LDS			FDS		
	MSE	L1(MAE)	G-Mean	MSE	L1(MAE)	G-Mean	MSE	L1(MAE)	G-Mean
Test Set (all Ages)	82.4131	6.3931	3.5687	156.1663	9.4174	5.8896	77.3760	6.1311	3.3490
Age < 18	66.8445	5.0118	2.2111	118.8197	6.8411	3.7566	52.5529	4.2850	1.9510
Age >= 80	326.3382	14.3916	9.5295	188.0271	10.4921	6.7884	266.016	12.5373	8.0696

From above chart, it is clear that three loss functions are consistent with pointing out the best model. Moreover, FDS is the best model when takes all age into account and when only takes age <18 into account. As we mentioned above, FDS transfers the feature statistics between nearby target bins so it allows borrowing of information between features to gain prediction power. However, LDS is the best model when only takes age>=80 into account. LDS allows borrowing of information between targets. For age>=80, observations are much less than other age ranges. LDS transfers statistics of age <80 to age >=80 to gain prediction power. However, for all age and age<18, where the observations are huge enough for prediction. This transfer of statistics will mix up distinct characteristics of different ages and make ages unidentifiable. Therefore, LDS has the worst performance on all age and age<18.

5 Concluding remarks

This project shows the advantages of LDS and FDS on deep imbalanced regression and also some limitation of LDS. LDS should be applied on datasets that have really sparse targets and huge data vacancy between targets. FDS is proved to be better than simple deep regression under any age range. Moreover, We note that FDS can be integrated with any neural network model, as well as any past work on improving label imbalance, even LDS.

References

SUBEDI, S. (2018, November). UTKFace, Version 1. Retrieved November 30, 2022 from https://www.kaggle.com/datasets/jangedoo/utkface-new

Yang, Y., Zha, K., Chen, Y.-C., Wang, H., & Katabi, D. (n.d.). *Delving into deep imbalanced regression*. – arXiv Vanity. Retrieved November 30, 2022, from https://www.arxiv-vanity.com/papers/2102.09554/