Key Papers

Survey on deep learning with class imbalance **(2019)**
Learning from imbalanced data: open challenges and future directions - Progress in Artificial Intelligence **(2016)**

Metrics for evaluation

Source	Content
SMOTE	In the presence of imbalanced datasets with unequal error costs, it is more appropriate to use the ROC curve or other similar techniques
Multi-class imbalanced big data classification on Spark	$AvAcc = \sum_{i=1}^{C} \frac{tp_i + tn_i}{tp_i + tn_i + fp_i + fn_i}$ $Rec_M = \frac{1}{C} \sum_{i=1}^{C} recall_i$ $Prec_M = \frac{1}{C} \sum_{i=1}^{C} precision_i$ $Rec_U = \sum_{i=1}^{C} tp_i \Big/ \sum_{i=1}^{C} t_i$ 14 $Prec_U = \sum_{i=1}^{C} tp_i \Big/ \sum_{i=1}^{C} p_i$ $F_{\beta M} = \frac{(1+\beta)^2 \cdot Prec_M \cdot Rec_M}{\beta^2 \cdot Prec_M + Rec_M}$ $F_{\beta\mu} = \frac{(1+\beta)^2 \cdot Prec_\mu \cdot Rec_\mu}{\beta^2 \cdot Prec_\mu + Rec_\mu}$ $AvF_\beta = \frac{1}{C} \sum_{i=1}^{C} \frac{(1+\beta^2) \cdot precision_i \cdot recall_i}{\beta^2 \cdot precision_i + recall_i}$ $CBA = \sum_{i=1}^{C} \frac{mat_{i,i}}{max(\sum_{j=1}^{C} mat_{i,j}, \sum_{j=1}^{C} mat_{j,i})}$

ODOC-ELM: Optimal decision outputs compensation-based extreme learning machine for classifying imbalanced data	$$\mathrm{F-measure} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$ $$\mathrm{G-mean} = \sqrt{\mathrm{TPR} \times \mathrm{TNR}}$$ where Precision, Recall, TPR and TNR are further defined as: $$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$ $$\mathrm{Recall} = \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$ $$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$
Boosting methods for multi-class imbalanced data classification: an experimental review	area under the curve (AUC), Matthews correlation coefficient (MCC), G-mean, Kappa, and others that some of them have been successfully extended to multi-class problems [19, 20].

Dataset / distribution used

What to find:

1. scenarios where there will be large amounts of data BUT highly imbalanced, binary or multiclass both ok. Need to be PICTURES!
2. What kind of distribution (long tailed) are these scenarios
3. What substitute datasets can we use to model these problems, or how do we modfiy existing datasets

Real life scenarios

Source	Content

<table>
<tr>
<td>Learning from imbalanced data: open challenges and future directions - Progress in Artificial Intelligence</td>
<td>
Table 1 A list of selected recent real-life applications with data imbalance present
<table>
<tr><th>Application area</th><th>Problem description</th></tr>
<tr><td>Activity recognition [19]</td><td>Detection of rare or less-frequent activities (multi-class problem)</td></tr>
<tr><td>Behavior analysis [3]</td><td>Recognition of dangerous behavior (binary problem)</td></tr>
<tr><td>Cancer malignancy grading [30]</td><td>Analyzing the cancer severity (binary and multi-class problem)</td></tr>
<tr><td>Hyperspectral data analysis [50]</td><td>Classification of varying areas in multi-dimensional images (multi-class problem)</td></tr>
<tr><td>Industrial systems monitoring [44]</td><td>Fault detection in industrial machinery (binary problem)</td></tr>
<tr><td>Sentiment analysis [65]</td><td>Emotion and temper recognition in text (binary and multi-class problem)</td></tr>
<tr><td>Software defect prediction [48]</td><td>Recognition of errors in code blocks (binary problem)</td></tr>
<tr><td>Target detection [45]</td><td>Classification of specified targets appearing with varied frequency (multi-class problem)</td></tr>
<tr><td>Text mining [39]</td><td>Detecting relations in literature (binary problem)</td></tr>
<tr><td>Video mining [20]</td><td>Recognizing objects and actions in video sequences (binary and multi-class problem)</td></tr>
</table>
</td>
</tr>
<tr>
<td>Boosting methods for multi-class imbalanced data classification: an experimental review</td>
<td>fault prediction [3], fraud detection [4], medical diagnosis [5], text classifcation [6], oil-spill detection in satellite images [7] and cultural modeling [8].</td>
</tr>
<tr>
<td>ODOC-ELM: Optimal decision outputs compensation-based extreme learning machine for classifying imbalanced data</td>
<td>biomedical applications [16] to network intrusion detection [17].</td>
</tr>
<tr>
<td>Survey on deep learning with class imbalance</td>
<td>Skewed data distributions naturally arise in many applications where the positive class occurs with reduced frequency, including data found in disease diagnosis [3], fraud detection [4, 5], computer security [6], and image recognition [7]. Intrinsic imbalance is the result of naturally occurring frequencies of data, e.g. medical diagnoses where the majority of patients are healthy. Extrinsic imbalance, on the other hand, is introduced through external factors, e.g. collection or storage procedures [28].</td>
</tr>
</table>

Datasets

Matrix factorisation

Source	Content
Survey on deep learning with class imbalance	**CIFAR-10 variation** Hensman and Masko [79] explored the effects of class imbalance and ROS using deep CNNs. The CIFAR-10 [80] benchmark data set, comprising 10 classes with 6000 images per class, was used to generate 10 imbalanced data sets for testing. Tese 10 generated data sets contained varying class sizes, ranging between 6% and 15% of the total data set, producing a max imbalance ratio ρ = 2.3. In addition to varying the class size, the different distributions also varied the number of minority classes, where a minority class is any class smaller than the largest class. For example, a major 50–50 split (Dist. 3) reduced fve of the classes to 6% of the data set size and increased fve of the classes to 14%. As another example, a major singular over-representation (Dist. 5) increased the size of the airplane class to 14.5%, reducing the other nine classes slightly to 9.5% **WHOI-Plankton** Lee et al. [20] combined RUS with transfer learning to classify highly-imbalanced data sets of plankton images, WHOI-Plankton [81]. The data set contains 3.4 million images spread over 103 classes, with 90% of the images comprising just five classes and the 5th largest class making up just 1.3% of the entire data set. Imbalance ratios of ρ > 650 are exhibited in the data set, with many classes making up less than 0.1% of the data set. **Self-collected dataset** The author’s self-collected data set contains over 10,000 images captured from publicly available network cameras, including a total of 19 semantic concepts, e.g. intersection, forest, farm, sky, water, playground, and park. From the original data set, 70% is used for training models, 20% is used for validation, and 10% is set aside for testing. The authors report imbalance ratios in the data set as high as ρ = 500 **MNIST, CIFAR-10, ImageNet Variation** Buda et al. [23] compare ROS, RUS, and two-phase learning using three multiclass image data sets and deep CNNs. MNIST [86], CIFAR-10, and ImageNet data sets are used to create distributions with varying levels of imbalance. Both MNIST and CIFAR-10 training sets contain 50,000 images spread evenly across 10 classes, i.e. 5000 images per class. **Imbalanced distributions were created from MNIST and CIFAR-10 in the range of ρ ∈ [10, 5000] and ρ ∈ [2, 50], respectively.** The ImageNet training data, containing 100 classes with a

	maximum of 1000 samples per class, was used to create imbalanced distributions in the range of $\rho \in [10, 100]$. **CIFAR-10. 20NewsGroups** A total of eight imbalanced binary data sets, including three image data sets and fve text data sets, were generated from the CIFAR100 [93] and 20 Newsgroup [94] collections. Te data sets are all relatively small, with most training sets containing fewer than 2000 samples and the largest training set containing just 3500 samples **COCO-Net** Lin et al. [88] Te COCO [99] data set was used to evaluate the proposed model against its competitors. **Buildings** Nemoto et al. [103] later used the focal loss in another image classification task, the automated detection of rare building changes, e.g. new construction. The airborne building images are annotated with the labels: no change, new construction, rebuilding, demolished, repaint roofs, and laying solar panels. The training data contains 203,358 images in total, where 200,000 comprise the negative class, i.e. no change. The repaint roofs and laying solar panel classes contain just 326 and 222 images, respectively, yielding class imbalance ratios as high as $\rho = 900$ **MNIST, CIFAR-100, Caltech-101, MIT-67, DIL, MLC** Khan et al. [19] introduced an effective cost-sensitive deep learning procedure which jointly learns network weight parameters and class misclassification costs during training. The proposed method, CoSen CNN, is evaluated against six multi-class data sets with varying levels of imbalance: MNIST, CIFAR-100, Caltech-101 [104], MIT-67 [105], DIL [106], and MLC [107]. Class imbalance ratios of $\rho = 10$ are tested for the MNIST, CIFAR-100, Caltech-101, and MIT-67 data sets. Te DIL and MLC data sets have imbalance ratios of $\rho = 13$ and $\rho = 76$, respectively **MNIST, CIFAR-10** Buda et al. [23] experimented with adjusting CNN output thresholds to improve overall performance. **They used the MNIST and CIFAR-10 data sets with varying levels of class imbalance ratios in the range of $\rho \in [1, 5000]$ and $\rho \in [1, 50]$, respectively** **CIFAR-10, CIFAR-100** Three class imbalanced distributions are created from each original data set through random under-sampling, i.e. Dist. A, Dist. B, and Dist. C. In Dist. A and Dist. B, half of the classes are reduced in size, creating imbalance levels of $\rho = 10$ and $\rho = 20$, respectively. In Dist. C, class

reduction levels increase linearly across all classes with a max imbalance of ρ = 20. For example, Dist. C for the CIFAR-100 data set contains 25 images in each of the frst 10 classes, 75 images in each of the next 10 classes, then 125 images in each of the next 10 classes, etc.

EmotioNet Challenge Track 1
The EmotioNet Challenge Track 1 data set [116] is used to compare methods. From the original data set, with over one million images, 450,000 images were randomly selected for training and 40,000 images were randomly selected for validation. Te images in this data set contain 11 possible FAUs. Since an image can be positive for more than one FAU, the authors treated the classifcation problem as a set of 11 binary problems. Several FAUs are present in less than 1% of the data set, e.g. nose wrinkler, chin raiser, and upper lid raiser. Te lip stretcher FAU is only present in 0.01% of the data, creating a max imbalance ratio to ρ = 10, 000.

CelebA
The proposed LMLE method is shown to achieve state-of-the-art results on the CelebA [119] data set, which contains high imbalance levels up to ρ = 49. The CelebA dataset contains facial images annotated with 40 attributes, with imbalance levels as high as ρ = 49 (Bald vs not Bald). A total of 160,000 images

MNIST, MNIST-back-rot, SVHN, CIFAR10, STL-10
Ando and Huang [117] introduced over-sampling to the deep feature space produced by CNNs in their DOS framework. The proposed method is extensively evaluated by generating imbalanced data sets from five popular image benchmark data sets, including MNIST, MNIST-back-rot, SVHN [124], CIFAR-10, and STL-10 [125].

CelebA, X-Domain, CIFAR-100
Experiments are conducted by extending state-of-the-art CNN architectures with the proposed method and performing classification on three benchmark data sets. The CelebA dataset contains a max imbalance level of ρ = 49. The X-Domain [127] data set contains 245,467 retail store clothing images that are annotated with 9 multi-class attribute labels, 165,467 of which are set aside for training. The X-Domain contains extreme class imbalance ratios of ρ > 4000. Several imbalanced data sets are generated from the CIFAR-100 set, with imbalance ratios up to ρ = 20, to demonstrate how CLR handles

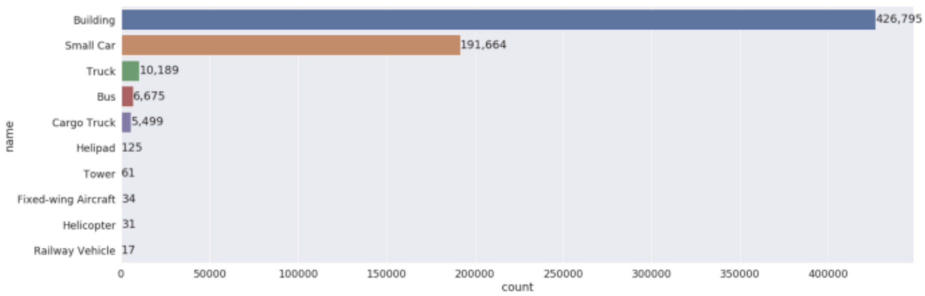
increasing levels of imbalance

Table 18 Summary of data sets and class imbalance levels

Paper	Data sets	Data type	Class count	Data set size	Min class size	Max class size	ρ (Eq. 1)
[79]	CIFAR-10	Image	10	60,000	2340	3900	2.3
[20]	WHOI-Plankton	Image	103	3,400,000	< 3500	2,300,000	657
[21]	Public cameras	Image	19	10,000	14	6986	499
[18]	CIFAR-100 (1)	Image	2	6000	150	3000	20
	CIFAR-100 (2)	Image	2	1200	30	600	20
	CIFAR-100 (3)	Image	2	1200	30	600	20
	20 News Group (1)	Text	2	1200	30	600	20
	20 News Group (2)	Text	2	1200	30	600	20
[88]	COCO	Image	2	115,000	10	100,000	10,000
[103]	Building changes	Image	6	203,358	222	200,000	900
[89]	GHW	Structured	2	2565	406	2159	5.3
	ORP	Structured	2	700	124	576	4.6
[19]	MNIST	Image	10	70,000	600	6000	10
	CIFAR-100	Image	100	60,000	60	600	10
	CALTECH-101	Image	102	9144	15	30	2
	MIT-67	Image	67	6700	10	100	10
	DIL	Image	10	1300	24	331	13
	MLC	Image	9	400,000	2600	196,900	76
[90]	KEEL	Structured	2	3339	26	3313	128
[91]	CIFAR-10	Image	10	60,000	250	5000	20
	CIFAR-100	Image	100	60,000	25	500	20
[22]	CelebA	Image	2	160,000	3200	156,800	49
[117]	MNIST	Image	10	60,000	50	5000	100
	MNIST-back-rot	Image	10	62,000	12	1200	100
	CIFAR-10	Image	10	60,000	5000	5000	1
	SVHN	Image	10	99,000	73	7300	100
	STL-10	Image	10	13,000	500	500	1
[118]	CelebA	Image	2	160,000	3200	156,800	49
[92]	EmotioNet	Image	2	450,000	45	449,955	10,000
[23]	MNIST	Image	10	60,000	1	5000	5000
	CIFAR-10	Image	10	60,000	100	5000	50
	ImageNet	Image	1000	1,050,000	10	1000	100

Images from CelebA and EmotioNet are treated as a set of binary classification problems, because they are each annotated with 40 and 11 binary attributes, respectively. The COCO data class imbalance arises from the extreme imbalance between background and foreground concepts

Multi-class imbalanced big data classification on Spark	uses 5 very large unbalanced datasets with varying number of classes, features and imbalance ratio, BUT none of them are image based. ● Covtype ● Traffic ● Seer ● Sensors ● Iot
ODOC-ELM: Optimal decision outputs compensation-based extreme learning machine for classifying imbalanced data	30 binary-class imbalanced data sets and 12 multiclass imbalanced data sets randomly acquired from the **Keel** data repository. All datasets are quite small. Unlikely to have need for dataset condensation.

<table>
<tr><td>SMOTE</td><td>
<table>
<tr><th>Dataset</th><th>Majority Class</th><th>Minority Class</th></tr>
<tr><td>Pima</td><td>500</td><td>268</td></tr>
<tr><td>Phoneme</td><td>3818</td><td>1586</td></tr>
<tr><td>Adult</td><td>37155</td><td>11687</td></tr>
<tr><td>E-state</td><td>46869</td><td>6351</td></tr>
<tr><td>Satimage</td><td>5809</td><td>626</td></tr>
<tr><td>Forest Cover</td><td>35754</td><td>2747</td></tr>
<tr><td>Oil</td><td>896</td><td>41</td></tr>
<tr><td>Mammography</td><td>10923</td><td>260</td></tr>
<tr><td>Can</td><td>435512</td><td>8360</td></tr>
</table>
Table 2: Dataset distribution
</td></tr>
<tr><td>https://towardsdatascience.com/4-ways-to-improve-class-imbalance-for-image-data-9adec8f390f1</td><td>


</td></tr>
</table>

Features of Datasets / Data distribution / Imbalance concepts

Source	Content
ODOC-ELM: Optimal decision outputs compensation-based extreme learning machine for classifying imbalanced data	In fact, the damage is related to multiple potential data distribution factors, including class overlap, imbalance ratio, the size of training instances, noisy data and small disjunctions [52,55–57].

http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=047D2F2C49B123F3AC37A700726A5D1A?doi=10.1.1.711.8214&rep=rep1&type=pdf (may be slightly outdated)	- Degree of concept complexity - Size of training set - Level of imbalance (imbalance ratio)

Math

Source	Content
Survey on deep learning with class imbalance	Anand et al. [78] explored the effects of class imbalance on the backpropagation algorithm in shallow neural networks in the 1990's. Te authors show that in class imbalanced scenarios, the length of the minority class's gradient component is much smaller than the length of the majority class's gradient component. In other words, the majority class is essentially dominating the net gradient that is responsible for updating the model's weights. Tis reduces the error of the majority group very quickly during early iterations, but often increases the error of the minority group and causes the network to get stuck in a slow convergence mode