Robust High-Dimensional Classification From Few Positive Examples

Deepayan Chakrabarti[1] , **Benjamin Fauber**[2]
[1]University of Texas, Austin
[2]Dell Inc.
deepay@utexas.edu, ben.fauber@dell.com

Abstract

We tackle an extreme form of imbalanced classification, with up to 10^5 features but as few as 5 samples from the minority class. This problem occurs in predicting predicting tumor types and fraud detection, among others. Standard imbalanced classification methods are not designed for such severe data scarcity. Sampling-based methods need too many samples due to the high-dimensionality, while cost-based methods must place too high a weight on the limited minority samples. Our proposed method, called DIRECT, bypasses sample generation by training the classifier over a robust smoothed distribution of the minority class. DIRECT is fast, simple, robust, parameter-free, and easy to interpret. We validate DIRECT on several real-world datasets spanning document, image, and medical classification. DIRECT is up to $5x - 7x$ better than SMOTE-like methods, $30-200\%$ better than ensemble methods, $3x - 7x$ better than cost-sensitive methods. The greatest gains are for settings with the fewest samples in the minority class, where DIRECT's robustness is most helpful.

1 Introduction

Suppose we have a corpus of articles tagged with topics, and we want to predict topic tags for new articles. Each article may have $O(10^5)$ keyword-based features. But there may be only a few articles on any one topic. For example, in a knowledge base of "help articles" at Dell Inc., half the topics had fewer than 9 articles (Figure 1). In such cases, a training set of articles for any single topic will have very few positive examples on that topic, but many negative examples. Similarly, to predict tumor types from genetic data, we may have around 10 positive examples (patients) but $O(10^4)$ features (genes) [Yeang *et al.*, 2001]. Acquiring much more data is infeasibly costly and time-consuming. Similar problems occur in fraud detection [Wei *et al.*, 2013] and cheminformatics [Czarnecki and Rataj, 2015]. For concreteness, consider the following problem which occurs in our experiments:

How do we train a classifier from 5 *positive points and* $1,000$ *negative points with* $100,000$ *features?*

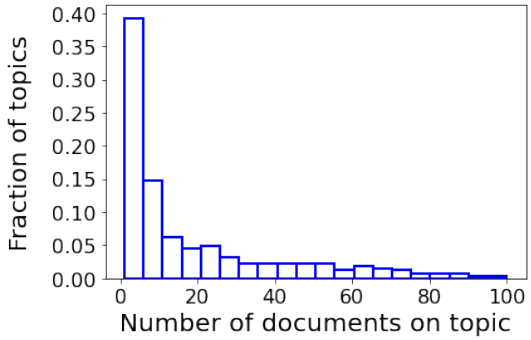


Figure 1: *Distribution of the number of documents per topic for a large knowledge base:* 50% of the topics had fewer than 9 documents per topic, and 70% had fewer than 26 documents per topic.

Existing methods mainly focus on the class imbalance. But, as we show empirically, the critical constraint is the extreme scarcity of minority class examples (e.g., only 5 samples in the question above). Sampling-based methods may generate biased samples when these are built from so few minority datapoints. Cost-sensitive methods may overfit to the minority class samples due to the overwhelming cost of misclassifying them. The same applies to complex classifiers. For example, deep-learning methods [Chung *et al.*, 2016] often need orders of magnitude more minority class samples. Simple methods outperform neural classifiers on limited training data [Cunha *et al.*, 2021; Yang *et al.*, 2019].

Our proposed method, called DIRECT,[1] offers a *robust and tractable* solution to the problem of *imbalanced binary classification under extreme data scarcity with no side information*. It is, to our knowledge, the first method to focus on this problem. Our approach is based on two ideas. First, we build a robust smooth distribution for the minority class, using only the statistics that can be reliably estimated. This robust distribution succinctly captures what can be inferred and what is uncertain about the true distribution. This makes DIRECT **robust to estimation errors**, which helps combat the problems of limited sample size. Note that this is different from adversarial robustness, since there is no data corruption involved and the training and test distributions are the same.

[1]Our code is available at https://github.com/deepayan12/direct.

Second, we show that DIRECT's expected loss under this robust distribution can be calculated in closed form. This enables DIRECT to bypass sample generation and train in a **single-step**. In effect, DIRECT optimizes over all possible samples, weighted by their probability. This improves both the speed and accuracy of the classifier.

DIRECT has several other appealing properties. With so few minority samples, cross-validation itself can be a source of error. However, DIRECT is **parameter-free**, and does not need cross-validation. DIRECT is also **fast and easily implementable** using any off-the-shelf optimizer.

Finally, we show **strong empirical results** on six text, two image, one medical, and twenty general UCI datasets. Most tasks have $10,000$ to $100,000$ features. DIRECT is up to $5x-7x$ better than sampling-based methods, $30-200\%$ better than ensemble methods, $3x-7x$ better than cost-sensitive SVMs. On most datasets and settings, DIRECT is either comparable or better than existing methods. The most significant gains occur in settings with the fewest minority class samples, where the robustness of DIRECT has the most impact.

The rest of the paper is organized as follows. We present details of DIRECT in Section 2. Empirical results are shown in Section 3. We survey the related literature in Section 4, and conclude in Section 5. Proofs and extra results are provided in the supplementary material.

2 Proposed Work

We are given an imbalanced training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}$, where each point i has a feature vector $\boldsymbol{x}_i \in \mathbb{R}^p$ and a class label $y \in \{+1, -1\}$. We assume, without loss of generality, that the positive class is the minority class. That is, if there are n_{lo} positive points and n_{hi} negative points in $\mathcal{D}$, then $n_{lo} \ll n_{hi}$. We focus on the setting where the number of features p is large but n_{lo} is small. For example, $p = 10^5$ and $n_{lo} \leq 7$ for the text classification example of Figure 1. We want a binary classifier with three properties:

- **(P1)** *It should be robust*, since estimation errors are unavoidable with small sample sizes.
- **(P2)** *It should not need sampling,* since we need many samples to model high-dimensional distributions, and such sampling takes time and makes results variable.
- **(P3)** *It should be parameter-free,* since tuning parameters by cross-validation is noisy with so few samples.

Our proposed classifier, called DIRECT (**D**istribution for **I**mbalanced data with **R**obust **E**stimation of **C**ovariance **T**echnique), achieves this by using a robust kernel density to model the minority class distribution. With an appropriate choice of loss function, we show that the expected loss under the robust density has a closed form. This simplifies the training of DIRECT. Next, we will discuss the robust density, the calculation of the expected loss, and the overall algorithm.

2.1 Robust Minority Class Distribution

To construct the robust distribution, we need statistics that we can reliably estimate even from a few samples. High-order moments are unreliable because they are sensitive to the distribution's tail, from which we may not have enough samples. Even the sample covariance is unreliable for small n_{lo} and high p. Formally, when n_{lo}/p is fixed but $n_{lo}, p \to \infty$, the sample covariance does not converge to the population covariance matrix [Marcenko and Pastur, 1967]. Hence, our robust distribution uses only the sample mean and a *robust* covariance estimate.

Our robust covariance estimators are shrinkage estimators [Ledoit and Wolf, 2004; Ledoit and Wolf, 2012]. Given a minority-class data matrix $X_{lo} \in \mathbb{R}^{n_{lo} \times p}$ with the singular value decomposition $X_{lo} = USV^T$, they estimate the covariance as $\Sigma_{lo} = VS'V^T + q \cdot I$. The diagonal matrix S' contains $min(n_{lo}, p)$ entries ($n_{lo} \ll p$ for us), and $q > 0$ is a scalar. *Both S' and q can be calculated optimally to minimize the expected estimation error of Σ_{lo}.* Hence, unlike regularization, we do not need cross-validation to find the best parameters. Note that we do not explicitly construct Σ_{lo}, which has $O(p^2)$ entries. We only calculate $V \in \mathbb{R}^{p \times min(n_{lo}, p)}$, S', and q.

Now, we use Σ_{lo} to construct a kernel density for the minority class. This places a smooth distribution (a "kernel") centered on each of the minority class points, and averages over them to get an overall density. Our choice for the kernel is against driven by robustness concerns. The kernel distribution should be the distribution with the greatest uncertainty (i.e., the maximum entropy) subject to zero mean (since it is centered on the points) and covariance Σ_{lo}. This is just the Gaussian distribution $\mathcal{N}(0, \Sigma_{lo})$ [Cover and Thomas, 2006]. Using the Gaussian kernel, our robust density for the minority class is given by:

$$p^\star_{lo}(\boldsymbol{x}) = \frac{1}{n_{lo}} \sum_{\{i|y_i=+1\}} \phi\left(\Sigma_{lo}^{-1/2}(\boldsymbol{x} - \boldsymbol{x}_i)\right) \tag{1}$$

$$\Sigma_{lo} = VS'V^T + q \cdot I, \tag{2}$$

where $\boldsymbol{x}_i$ is a point from the minority class ($y_i = +1$), $\phi(.)$ is the $\mathcal{N}(0, 1)$ density, and (V, S', q) are discussed above.

We note that kernel density estimates also have a bandwidth parameter. For large p and small n_{lo}, this bandwidth approaches 1 and can be ignored (see Eq. 4.14 and Table 4.1 of [Silverman, 1986]).

2.2 Model-Fitting

Consider a classifier with parameters $\boldsymbol{\theta}$ and loss function $\ell(y, \boldsymbol{x}; \boldsymbol{\theta})$. We must train it using n_{lo} points from the minority class and n_{hi} from the majority class, with $n_{lo} \ll n_{hi}$. For the majority class, n_{hi} is large enough that the average loss is a good proxy for the test loss. But for the minority class, n_{lo} is too small, and the average loss can be noisy. So we use the expected loss over the robust distribution (Eq. 1). Hence, we find the model parameters $\boldsymbol{\theta}$ to minimize

$$\frac{\sum_{\{i|y_i=-1\}} \ell(y_i = -1, \boldsymbol{x}_i; \boldsymbol{\theta})}{n_{hi}} + E_{\boldsymbol{z} \sim p^\star_{lo}} \ell(y_i = +1, \boldsymbol{z}; \boldsymbol{\theta}). \tag{3}$$

We can approximate the second term by averaging the loss over samples drawn from $p^\star_{lo}(.)$. However, we can avoid this sampling step by choosing a loss whose expectation has a closed form. In this paper, we choose $\boldsymbol{\theta} = (c, \boldsymbol{w}) \in \mathbb{R} \times \mathbb{R}^p$ and $\ell(y, \boldsymbol{x}; \boldsymbol{\theta} = (c, \boldsymbol{w})) = \max(0, 1 - y \cdot (c + \boldsymbol{w}^T \boldsymbol{x}))$. In

other words, given a feature vector $\boldsymbol{x}$, our model predicts $y = \text{sign}(c + \boldsymbol{w}^T\boldsymbol{x})$ and uses the hinge loss with unit margin to measure accuracy. This loss is convex in $\boldsymbol{\theta}$, so the objective of Eq. 3 is convex too. Furthermore, we can calculate the second term of Eq. 3 in closed form, as shown next (all proofs are deferred to the supplementary material).

Theorem 1.

$$\begin{aligned} E_{\boldsymbol{z}\sim p^\star_{lo}}\ell(y_i=+1,\boldsymbol{z};\boldsymbol{\theta}) \\ =&(1/n_{lo})\sum_{i|y_i=+1} s_i\cdot\Phi\left(\frac{s_i}{t}\right)+t\cdot\phi\left(\frac{s_i}{t}\right), \quad (4)\\ s_i =&1-(c+\boldsymbol{w}^T\boldsymbol{x}_i),\\ t =&\sqrt{\boldsymbol{w}^T\Sigma_{lo}\boldsymbol{w}}. \end{aligned}$$

Here, $\phi(\cdot)$ and $\Phi(\cdot)$ represent the pdf and cdf of a standard normal respectively.

For intuition, suppose $t \approx 0$. Then the expected loss for point i from the positive class becomes $s_i \cdot \Phi(s_i/t) + t \cdot \phi(s_i/t) \approx s_i \cdot \mathbb{1}_{s_i>0} = \max(0, s_i)$, which is exactly the hinge loss for i. Thus, for $t \approx 0$, we recover the empirical loss. As t increases, the loss increases, as shown below.

Corollary 1. *The expected loss over the robust distribution (Eq. 4) is an increasing function of both s_i and t.*

Thus, $t = \|\Sigma_{lo}^{1/2}\boldsymbol{w}\|$ acts as a regularizer for $\boldsymbol{w}$. But unlike standard regularization, this is not a norm of $\boldsymbol{w}$, and we do not need cross-validation to select the regularization penalty.

Using Theorem 1, Eq. 3 has a closed-form formula:

$$\begin{aligned}\min_{\boldsymbol{w}\in\mathbb{R}^p, c\in\mathbb{R}} \quad & \frac{1}{n_{hi}}\cdot\sum_{i|y_i=-1}\max\left(0,1+(c+\boldsymbol{w}^T\boldsymbol{x}_i)\right)\\ &+\frac{1}{n_{lo}}\cdot\sum_{i|y_i=+1}\left(s_i\cdot\Phi\left(\frac{s_i}{t}\right)+t\cdot\phi\left(\frac{s_i}{t}\right)\right) \quad (5)\end{aligned}$$

Eq. 5 is a convex problem, and we can use any off-the-shelf solver to minimize it. A significant speedup is provided by the following theorem.

Theorem 2. *Let $(\boldsymbol{w}_\star, c)$ be the minimizer of Eq. 5. Then $\boldsymbol{w}_\star$ lies in the subspace spanned by the $n_{lo}+n_{hi}$ vectors $\{\boldsymbol{x}_i\}$.*

Hence, when $n_{lo} + n_{hi} \ll p$ (as in our case), we can assume $\boldsymbol{w}_\star = \sum_i \alpha_i \boldsymbol{x}_i$ and only search for the optimal $\boldsymbol{\alpha} \in \mathbb{R}^{n_{lo}+n_{hi}}$. This reduces the search space for the convex solver from p dimensions to $n_{lo} + n_{hi}$ dimensions.

2.3 Post-Processing

The process above yields a feature vector $\boldsymbol{w}_\star$ and an intercept c to minimize the loss function of Eq. 3. However, our overall objective is classification accuracy over a balanced test set. So, in a post-processing step, we adjust the intercept for this objective. In particular, we minimize the expected misclassification:

$$\begin{aligned}\min_{c'\in\mathbb{R}} \frac{1}{n_{hi}}\cdot\sum_{i|y_i=-1}&\mathbb{1}_{c'+\boldsymbol{w}_\star^T\boldsymbol{x}_i}\\ +\frac{1}{n_{lo}}\cdot\sum_{i|y_i=+1}&\underbrace{E_{\boldsymbol{z}_i\sim\mathcal{N}(\boldsymbol{x}_i,\Sigma_{lo})}\mathbb{1}_{-(c'+\boldsymbol{w}_\star^T\boldsymbol{z}_i)}}_{=\Phi\left(-\frac{c'+\boldsymbol{w}_\star^T\boldsymbol{x}_i}{\sqrt{\boldsymbol{w}_\star^T\Sigma_{lo}\boldsymbol{w}_\star}}\right)}\cdot \quad (6)\end{aligned}$$

Algorithm 1 DIRECT

1: **function** DIRECT(Training set $X_{lo} \in \mathbb{R}^{n_{lo}\times p}, X_{hi} \in \mathbb{R}^{n_{hi}\times p}$)
2: $\quad\Sigma_{lo} \leftarrow$ robust covariance for minority class (Eq. 2)
3: $\quad(\boldsymbol{w}_\star, c) \leftarrow$ minimize the convex loss of Eq. 5
4: $\quad c_\star \leftarrow$ minimize misclassification loss of Eq. 6
5: $\quad$**return** $(\boldsymbol{w}_\star, c_\star)$
6: **end function**

Any off-the-shelf single-variable optimizer can be used to find the minimizer $c_\star$. Algorithm 1 shows the pseudocode for training DIRECT. This returns the optimal parameters $(\boldsymbol{w}_\star, c_\star)$. Now, given a test point $\boldsymbol{x}$, DIRECT predicts $\hat{y} = \text{sign}(c_\star + \boldsymbol{x}^T\boldsymbol{w}_\star)$.

Matching desired properties. DIRECT achieves robustness to limited data (property $(\mathbf{P1})$) by only using reliable statistics such as the optimal shrinkage covariance. The resulting robust distribution (Eq. 1) reflects these statistics, but also the uncertainty about other moments. DIRECT avoids sampling (property $(\mathbf{P2})$) by using a closed-form formula for the expected loss under the robust distribution (Theorem 1). This speeds up model-fitting and also avoids the variability of sampling. Finally, DIRECT's only parameters are the shrinkage parameters for the robust covariance Σ_{lo} (Eq. 2). We can compute the optimal shrinkage from the data itself, without cross-validation. Thus, we achieve property $(\mathbf{P3})$.

Remark 1. *We note that any complex classifier can be trained within our robust framework via sampling. In other words, we can approximate the expected loss (Eq. 3) by sampling from the robust distribution $p^\star_{lo}$, and use this approximation for parameter-fitting. This retains the robustness to limited data, but loses the speed of* DIRECT. *Another alternative is available if we want a classifier based on functions of $\boldsymbol{x}$ (say, a finite number of interaction terms). Then, we can construct new feature vectors $\tilde{\boldsymbol{x}}_i$ from functions of $\boldsymbol{x}_i$, and use* DIRECT *on this modified dataset.*

3 Experiments

We will discuss three questions: (a) how accurate is DIRECT, (b) how does the accuracy vary with the imbalance ratio and the minority class sample size, and (c) how fast is DIRECT?

Experimental setup. We ran experiments on six text, two image, and one medical dataset, along with 20 UCI datasets (Table 1). In each experiment, we created a training set with n_{lo} positive and n_{hi} negative samples that were randomly chosen from the dataset. All remaining datapoints were used for testing. We ran experiments on 509 unique (dataset, class, n_{lo}, n_{hi}) combinations, each being repeated 30 times.

Note that *our focus is on the limited data binary classification problem* (as motivated by Figure 1). So, we used $n_{lo} \leq 50$ for almost all experiments. Also, we do not consider multiclass classification or specialized feature construction methods. Multiclass learning can be implemented via repeated one-versus-all classification using DIRECT. Specialized features may not be available (e.g., for our medical

Type	Dataset	Features	Classes
Text	Dell	100,000	8
	20-Newsgroups	25,804	20
	Reuters	20,000	21
	News Search	20,000	11
	Arxiv	20,000	10
	Recipes	10,000	8
Medical	Tumors	16,063	15
Image	MNIST (digits)	784	10
	MNIST (fashion)	784	10
Misc.	UCI (20 datasets)	8 . . . 561	20

Table 1: Dataset statistics.

dataset); when available, they can be used as inputs to DIRECT. Hence, we do not pursue these topics in this paper. The supplementary material has more details about our datasets.

Accuracy metric. Our comparison metric is the *area under the precision-recall curve (AUPRC)*, which is the standard accuracy metric for imbalanced binary classification [Davis and Goadrich, 2006]. The AUPRC scores for all 509 experiments are presented in the supplementary material. In this section, we report the lift in AUPRC of DIRECT over competing methods. The lift provides an aggregate measure of the outperformance of DIRECT across all classes for a dataset. The supplementary material also shows alternate metrics such as the balanced accuracy and the G-mean, which show similar patterns as the AUPRC results.

Competing methods. Among data balancing methods, we consider SMOTE [Chawla *et al.*, 2002], ADASYN [He *et al.*, 2008], Borderline-SMOTE [Han *et al.*, 2005], and oversampling with smoothed bootstraps (ROSE) [Menardi and Torelli, 2014]. These are coupled with a linear SVM so that the number of parameters is the same as in DIRECT. Among ensemble methods, we show results using balanced random forests, balanced boosting (RUSBoost), SMOTE with Gradient Boosting (XGBoost), and balanced bagged decision trees [Chen *et al.*, 2004; Seiffert *et al.*, 2010]. We also show results against a cost-sensitive SVM, an imbalance-aware margin classifier (LDAM-RDW) [Cao *et al.*, 2019], and vanilla SVM (SVC).

Cost-sensitive deep learning [Chung *et al.*, 2016] underperformed other baselines even after tuning hyperparameters on the test sets, so we do not show those results. [Cunha *et al.*, 2021] make similar observations about deep learning for limited data. Our code will be made publicly available after incorporating comments from the reviewers.

3.1 Accuracy Comparison

Table 2 shows the trimmed mean of DIRECT's lift in AUPRC. The trimmed statistics are robust to noise due to the limited sample sizes. Detailed results for all 509 problem settings are in the supplementary material.

We see that **DIRECT is $5\mathbf{x} - 7\mathbf{x}$ better than all SMOTE-like methods on the Tumors dataset.** For example, with $n_{lo} = 5$, the median AUPRC over all classes is 0.59 for DIRECT but only 0.07 (close to random) for the SMOTE-based methods. On other datasets, with $n_{lo} \leq 20$, DIRECT typically outperforms SMOTE-like methods. When $n_{lo} = 50$, DIRECT is better for News Search and the two MNIST datasets. Overall, DIRECT outperforms SMOTE-based methods significantly on several datasets and is comparable otherwise.

DIRECT is $77 - 145\%$ better than bagging, $50 - 74\%$ better than random forests, and $30 - 210\%$ better than boosting for certain settings. The most significant differences are for Arxiv, MNIST (digits), Recipes, and Tumors. The closest competitor is SMOTE with gradient boosting, which is better than DIRECT for 20Newsgroups but is comparable or significantly worse on other datasets (e.g., DIRECT is $180\% - 215\%$ better on Dell).

DIRECT is $3\mathbf{x} - 7\mathbf{x}$ better than cost-sensitive SVMs and $75 - 80\%$ better than ROSE on the Tumors dataset. The closest competitor of DIRECT is ROSE-SVM. But, in most cases, DIRECT is comparable or better than it. DIRECT outperforms LDAM-RDW, showing that for small n_{lo}, the robustness to limited data is more important than margin adaptation. DIRECT also outperforms SVC by a wide margin.

Figure 2 shows the distribution of the lift, instead of just the trimmed mean. It confirms that DIRECT can be much better than other methods but is rarely much worse. The most significant outperformance is for small n_{lo}, where the robustness of DIRECT has the most impact.

3.2 Varying the Training Size

In practical applications such as document tagging (Figure 1), the minority class sample size n_{lo} is often very limited. Our previous experiments focused on such settings. DIRECT's success resulted from its robustness to limited data. Now, we investigate further the relative impact of limited data and the imbalance ratio on DIRECT's accuracy.

Figure 3a shows how DIRECT's AUPRC varies with the size of the training data for the News Search dataset. When $n_{lo} = 5$, the AUPRC increases only 38% even as n_{hi} increases by a factor of 40. Thus, AUPRC improves in spite of increasing imbalance. But when $n_{hi} = 1000$, the AUPRC increases quickly with n_{lo} before flattening at $n_{lo} = 100$. Thus, increasing n_{lo} from 5 to 100 provides a significant benefit, but further reduction in imbalance does not help. The greatest gains occur when both n_{lo} and n_{hi} increase in lockstep.

In our problem setting, the minority class has extreme data scarcity. Here, these results suggest that *the limited sample size is the main problem, and not the imbalance.* Even a few more minority class samples helps accuracy, but more majority class samples have low marginal benefit. This shows the utility of DIRECT's robustness to limited data.

3.3 Wall-Clock Time

Figure 3b show the wall-clock time for training on the News Search dataset. We observe:

- By avoiding sampling, DIRECT is significantly faster than ROSE and SMOTE with XGBoost, both of which generate new samples for the minority class.
- By being parameter-free, DIRECT is faster than even SVC, which needs cross-validation for choosing the regularization parameter.

DIRECT also outperforms both in terms of accuracy. In testing, DIRECT was 30x faster than balanced bagging and boosting classifiers. This is because DIRECT only calculates a lin-

Lift in AUPRC of DIRECT over		SMOTE	Borderline SMOTE	ADASYN	ROSE	Balanced Decision Tree	Balanced Random Forest	SMOTE + Grad. Boost	Balanced Boosting	Cost sensitive SVM	LDAM-DRW	SVC
n_{lo}	n_{hi}											
		Tumors										
100	3	x	x	x	77.5%*	97.3%*	52.8%*	x	74.3%*	300.7%*	411.0%*	284.4%*
100	5	474.2%*	474.2%*	474.8%*	78.8%*	101.7%*	51.1%*	33.6%*	77.6%*	474.2%*	623.2%*	434.3%*
100	7	698.2%*	698.3%*	698.3%*	83.7%*	95.8%*	55.5%*	33.8%*	63.7%*	698.3%*	969.7%*	607.8%*
		News Search										
1000	5	6.3%	6.3%	6.4%	1.3%	24.4%*	8.8%*	10.9%*	29.4%*	6.4%	29.5%*	51.7%*
1000	10	7.3%	7.1%	7.3%	2.1%	33.1%*	9.5%*	18.4%*	44.1%*	7.3%	24.7%*	73.8%*
1000	20	8.1%	7.9%	8.1%	4.2%	37.5%*	9.6%*	16.7%*	55.9%*	8.1%	17.1%*	98.3%*
1000	50	6.5%	6.3%	6.6%	5.4%	29.8%*	5.9%	15.1%*	73.3%*	6.5%	10.1%*	58.0%*
		Arxiv										
1000	5	21.2%*	20.8%*	21.2%*	6.3%	143.1%*	73.8%*	78.7%*	187.4%*	21.2%*	105.8%*	583.9%*
1000	10	14.5%*	13.8%*	14.5%*	3.9%	125.1%*	60.9%*	61.9%*	195.0%*	14.5%*	62.5%*	776.8%*
1000	20	6.1%*	5.9%*	6.1%*	0.2%	95.9%*	42.0%*	41.7%*	195.1%*	6.1%*	29.9%*	888.1%*
1000	50	-1.3%	-1.4%	-1.3%	-2.1%	58.5%*	27.9%*	31.6%*	209.1%*	-1.3%	12.1%*	25.7%
		MNIST (digits)										
1000	5	68.7%*	68.7%*	68.7%*	22.0%*	76.9%*	14.3%*	36.4%*	34.6%*	68.7%*	555.4%*	130.0%*
1000	10	61.9%*	61.9%*	62.1%*	19.2%*	42.4%*	6.7%*	7.4%*	33.2%*	61.9%*	628.9%*	126.7%*
1000	20	51.0%*	51.0%*	51.2%*	16.0%*	23.2%*	1.5%	0.1%	24.5%*	51.0%*	715.9%*	94.2%*
1000	50	38.7%*	38.7%*	38.7%*	13.1%*	7.4%*	-3.1%*	-4.3%*	27.9%*	38.7%*	615.5%*	41.5%*
		MNIST (fashion)										
1000	5	36.4%*	36.5%*	36.4%*	1.5%	38.8%*	10.0%*	29.4%*	20.1%*	36.4%*	525.3%*	56.8%*
1000	10	34.3%*	34.3%*	34.3%*	2.0%	21.8%*	6.1%*	7.4%*	18.1%*	34.3%*	577.6%*	53.1%*
1000	20	28.8%*	28.8%*	28.8%*	1.0%	12.2%*	2.6%	-0.1%	16.4%*	28.8%*	612.5%*	33.3%*
1000	50	20.9%*	20.8%*	20.8%*	0.1%	3.3%*	-2.2%	-4.2%	17.6%*	20.8%*	512.8%*	8.7%*
		Reuters										
1000	5	11.7%	11.6%	11.7%	-0.2%	20.5%*	37.8%*	22.0%*	13.1%	11.7%	52.7%*	498.6%*
1000	10	6.1%	5.7%	6.1%	-0.1%	3.9%	22.8%*	8.0%*	14.1%*	6.1%	24.4%*	297.4%*
1000	20	2.6%	2.4%	2.6%	-0.9%	-2.8%	12.5%*	0.4%	13.0%*	2.6%	5.0%	63.9%*
1000	50	-0.2%	-0.4%	-0.2%	-1.9%	-13.1%*	7.1%	-6.1%	13.3%*	-0.2%	-5.9%	0.8%
		20-Newsgroups										
1000	5	41.1%*	41.1%*	41.1%*	32.9%*	43.4%*	26.2%*	-19.5%*	63.0%*	41.1%*	138.5%*	127.7%*
1000	10	35.3%*	35.3%*	35.3%*	21.3%*	59.2%*	39.5%*	-16.2%*	107.9%*	35.3%*	250.7%*	238.4%*
1000	20	8.4%*	8.3%*	8.3%*	-7.1%*	60.3%*	41.7%*	-15.4%*	147.0%*	8.3%*	431.4%*	255.4%*
1000	50	-0.8%	-0.8%	-0.8%	-12.5%*	38.1%*	30.9%*	-6.3%*	164.9%*	-0.8%	684.9%*	1.3%
		Recipes										
1000	5	0.1%	-0.5%	0.1%	-0.4%	78.7%*	66.1%*	30.5%*	77.9%*	0.1%	52.4%*	154.5%*
1000	10	-0.1%	-0.6%	-0.1%	0.4%	64.2%*	62.9%*	24.1%*	89.1%*	-0.1%	38.5%*	187.4%*
1000	20	-0.1%	-0.4%	-0.1%	2.2%	53.4%*	49.0%*	17.5%*	92.5%*	-0.1%	20.0%*	124.3%*
1000	50	-0.1%	-0.1%	-0.1%	3.0%	35.7%*	31.4%*	15.9%*	102.9%*	-0.1%	9.0%*	0.1%
		Dell										
1000	3	x	x	x	0.7%	x	42.8%*	x	65.4%*	2.1%	467.6%*	943.0%*
1000	5	0.5%	0.5%	0.5%	0.4%	42.2%*	37.2%*	214.5%*	71.9%*	0.5%	190.6%*	627.4%*
1000	7	-0.2%	-0.2%	-0.2%	0.1%	27.1%*	27.0%*	182.9%*	55.3%*	-0.2%	170.3%*	239.8%*
		UCI										
100	5	1.5%	1.7%	1.5%	3.0%*	4.4%*	2.0%	7.3%*	6.0%*	1.8%	30.1%*	5.9%*
100	20	1.4%	1.6%	1.7%	5.2%*	3.3%*	0.1%	2.9%*	8.2%*	1.6%	42.7%*	2.6%
100	50	1.7%	2.0%*	1.9%*	8.2%*	1.6%*	-0.8%	1.7%*	6.6%*	1.7%*	47.8%*	1.0%
1000	100	1.6%	1.9%	1.9%*	6.4%*	1.5%*	-1.6%	-0.6%	11.9%*	1.4%	64.0%*	4.7%*

Table 2: *AUPRC comparison:* The table shows the trimmed-mean (over 30 experiments) of the lift of DIRECT over other methods. Higher the lift, greater the outperformance of DIRECT. Stars indicate statistical significance ($p < 0.01$), and methods that fail to complete are marked with a cross. The robustness of DIRECT enables it to outperform other methods in head-to-head comparisons, particularly for small n_{lo}.

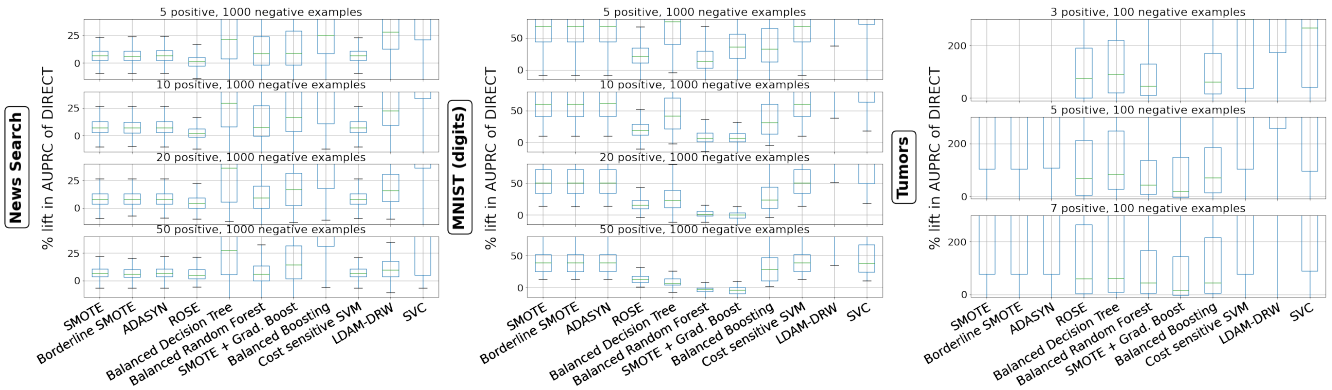
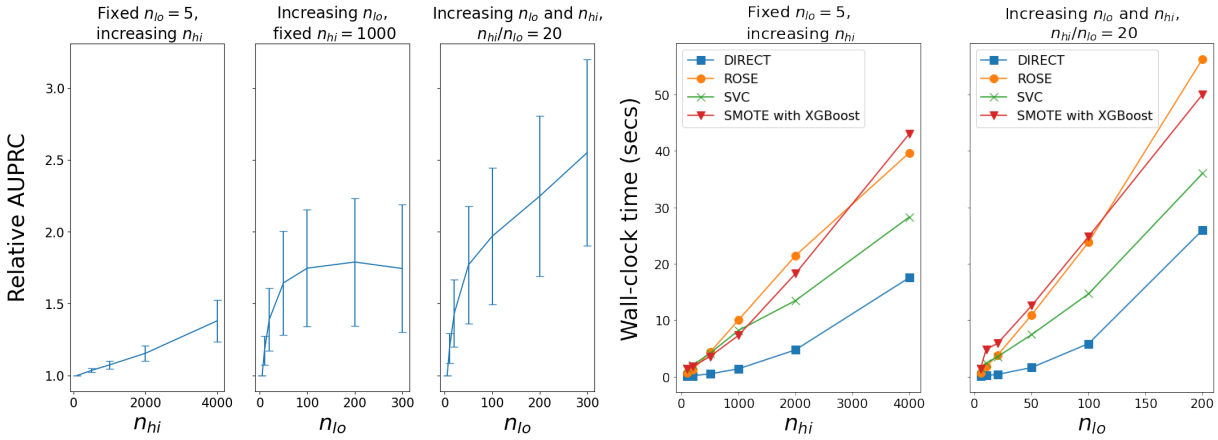


Figure 2: Distribution of the lift in AUPRC of DIRECT (higher is better).



(a) Change in AUPRC with varying sample size.

(b) Wall-clock times with varying sample size.

Figure 3: Effect of sample sizes on AUPRC and wall-clock time.

ear function for a given test point, while ensemble methods need more complex computations.

4 Related Work

We can divide prior work on imbalanced classification into three groups: data modification, ensemble methods, and other techniques. We also discuss robust approaches.

Data modification. SMOTE and its variants create a balanced dataset by generating synthetic samples [Chawla *et al.*, 2002; Han *et al.*, 2005; He *et al.*, 2008]. However, SMOTE's synthetic points always lie within the convex hull of the minority class points. For limited data, these may only lie in a subspace of the feature space. Also, nearest neighbors are costly to compute and lose their intuitive meaning in high dimensions.

ROSE generates samples using a Gaussian kernel with a diagonal covariance [Menardi and Torelli, 2014]. Unlike DIRECT, a diagonal covariance ignores feature correlations. Further, given limited training data, the diagonal covariance may be noisy. Finally, ROSE needs sampling from the high-dimensional kernel, unlike DIRECT.

Ensemble methods. These methods combine results from multiple models trained on balanced versions of the imbalanced data. On these balanced datasets, one can train random forests, boosted decision trees, and others [Chen *et al.*, 2004; Seiffert *et al.*, 2010]. But when the training data is limited but the number of features is high, ensemble methods can underperform, as we show empirically.

Cost-sensitive methods. There is work on modifying existing algorithms to allow for different misclassification costs [Chung *et al.*, 2016] or class-specific margins [Cao *et al.*, 2019]. However, cost-sensitive methods (including deep learning) underperform for limited datasets [Cunha *et al.*, 2021]. Recent works also consider semi-supervised imbalanced classification [Yang and Xu, 2020] or use extra supervision [Meng *et al.*, 2018]. We leave the extension of DIRECT to these settings for future work.

Robust algorithms. Robust methods consider uncertainty sets for uncertain or perturbed data or data distributions [Xu *et al.*, 2009; Tzelepis *et al.*, 2018; Mohajerin Esfahani and Kuhn, 2018]. But none of them consider data as limited as ours. Chakrabarti [2021] provides a theoretical justification for similar robust classifiers, but does not account for class imbalance.

5 Conclusions

We considered binary classification problems with imbalanced classes, few samples from the minority class, and high dimensionality. We devised a new algorithm called DIRECT for this extreme setting. DIRECT is simple, parameter-free, and robust to estimation error. In contrast to many existing methods, DIRECT does not need to generate samples. Instead, DIRECT incorporates a smooth estimate of the minority class distribution directly in its loss function. The estimated distribution accounts for correlated features and is robust to estimation error. DIRECT's loss function is convex and easily optimized via off-the-shelf solvers.

We empirically validated DIRECT on several real-world classification tasks on document, image, and gene microarray datasets. DIRECT is often significantly better than existing methods and rarely worse. DIRECT is up to $5x - 7x$ better than SMOTE-like methods, $30 - 200\%$ better than ensemble methods, $3x - 7x$ better than cost-sensitive SVMs. The most significant improvements often occur for the smallest sample sizes. That is because DIRECT's robustness to estimation error is most helpful in these settings.

One direction for future work is to extend DIRECT from binary to multiclass and multilabel classification. Since individual one-versus-all classifiers may have different accuracies, we need to calibrate them before combining them. Another is to leverage sparse feature vectors, such as in document classification. Finally, one can extend DIRECT to complex classifiers, e.g., by sampling from the smoothed distribution. This may be useful when there is enough data for the minority class.

References

[Cao *et al.*, 2019] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In *NeuRIPS*, 2019.

[Chakrabarti, 2021] Deepayan Chakrabarti. Robust linear classification from limited training data. *Machine Learning*, 2021.

[Chawla *et al.*, 2002] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002.

[Chen *et al.*, 2004] Chao Chen, Andy Liao, and Leo Breiman. Using Random Forest to Learn Imbalanced Data. Technical Report 666, UC Berkeley, 2004.

[Chung *et al.*, 2016] Yu-An Chung, Hsuan-Tien Lin, and Shao-Wen Yang. Cost-aware pre-training for multiclass cost-sensitive deep learning. In *IJCAI*, 2016.

[Cover and Thomas, 2006] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006.

[Cunha *et al.*, 2021] Washington Cunha, Vítor Mangaravite, Christian Gomes, Sérgio Canuto, Elaine Resende, Cecilia Nascimento, Felipe Viegas, Celso França, Wellington Santos Martins, Jussara M. Almeida, Thierson Rosa, Leonardo Rocha, and Marcos André Gonçalves. On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study. *Information Processing & Management*, 58(3):102481, May 2021.

[Czarnecki and Rataj, 2015] W. M. Czarnecki and K. Rataj. Compounds Activity Prediction in Large Imbalanced Datasets with Substructural Relations Fingerprint and EEM. In *2015 IEEE Trustcom/BigDataSE/ISPA*, volume 2, pages 192–192, August 2015.

[Davis and Goadrich, 2006] Jesse Davis and Mark Goadrich. The relationship between Precision-Recall and ROC curves. In *ICML*, pages 233–240, 2006.

[Han *et al.*, 2005] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In *International Conference on Intelligent Computing*, pages 878–887, 2005.

[He *et al.*, 2008] Haibo He, Yang Bai, E. A. Garcia, and Shutao Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *International Joint Conference on Neural Networks*, pages 1322–1328, 2008.

[Ledoit and Wolf, 2004] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, February 2004.

[Ledoit and Wolf, 2012] Olivier Ledoit and Michael Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060, April 2012.

[Marcenko and Pastur, 1967] V. A. Marcenko and L. A. Pastur. Distribution of Eigenvalues for Some Sets of Random Matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

[Menardi and Torelli, 2014] Giovanna Menardi and Nicola Torelli. Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28(1):92–122, January 2014.

[Meng *et al.*, 2018] Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. Weakly-Supervised Neural Text Classification. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 983–992, Torino Italy, October 2018. ACM.

[Mohajerin Esfahani and Kuhn, 2018] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, September 2018.

[Seiffert *et al.*, 2010] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. 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, January 2010.

[Silverman, 1986] B. W. Silverman. *Density estimation for statistics and data analysis*. Number 26 in Monographs on statistics and applied probability. Chapman & Hall/CRC, London, 1986.

[Tzelepis *et al.*, 2018] Christos Tzelepis, Vasileios Mezaris, and Ioannis Patras. Linear Maximum Margin Classifier for Learning from Uncertain Data. *IEEE PAMI*, 40(12):2948–2962, December 2018.

[Wei *et al.*, 2013] Wei Wei, Jinjiu Li, Longbing Cao, Yuming Ou, and Jiahang Chen. Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web*, 16(4):449–475, 2013.

[Xu *et al.*, 2009] Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and Regularization of Support Vector Machines. *Journal of Machine Learning Research*, 10:1485–1510, 2009.

[Yang and Xu, 2020] Yuzhe Yang and Zhi Xu. Rethinking the Value of Labels for Improving Class-Imbalanced Learning. In *NeurIPS*, page 12, 2020.

[Yang *et al.*, 2019] Wei Yang, Kuang Lu, Peilin Yang, and Jimmy Lin. Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models. In *SIGIR*, 2019.

[Yeang *et al.*, 2001] C. H. Yeang, S. Ramaswamy, P. Tamayo, S. Mukherjee, R. M. Rifkin, M. Angelo, M. Reich, E. Lander, J. Mesirov, and T. Golub. Molecular classification of multiple tumor types. *Bioinformatics (Oxford, England)*, 17 Suppl 1:S316–322, 2001.