Robust High-Dimensional Classification From Few Positive Examples

Deepayan Chakrabarti\textsuperscript{1}, Benjamin Fauber\textsuperscript{2}

\textsuperscript{1}University of Texas, Austin
\textsuperscript{2}Dell Inc.

deejay@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 $O(10^5)$ 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:

\textit{How do we train a classifier from 5 positive points and 1,000 negative points with 100,000 features?}

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,\textsuperscript{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.

\textsuperscript{1}Our code is available at https://github.com/deepayan12/direct.

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.
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 $D = \{(x_i, y_i)\}$, where each point $i$ has a feature vector $x_i \in \mathbb{R}^p$ and a class label $y_i \in \{+1, -1\}$. 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 $D$, 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 (Distribution for Imbalanced data with Robust Estimation of Covariance Technique), 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 $X_{lo} = U S V^T$, they estimate the covariance as $\Sigma_{lo} = V S' V^T + q \cdot I$. 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 $\Sigma_{lo}$. Hence, unlike regularization, we do not need cross-validation to find the best parameters.

Now, we use $\Sigma_{lo}$ 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 $\Sigma_{lo}$. 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:

$$p_{lo}^*(x) = \frac{1}{n_{lo}} \sum_{(i | y_i = +1)} \phi \left( \Sigma_{lo}^{-1/2} (x - x_i) \right)$$

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

where $x_i$ 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 $\theta$ and loss function $\ell(y, x; \theta)$. 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 $\theta$ to minimize

$$\frac{1}{n_{hi}} \sum_{(i | y_i = -1)} \ell(y_i = -1, x_i; \theta) + E_{z \sim p_{lo}^*} \ell(y_i = +1, z; \theta).$$

We can approximate the second term by averaging the loss over samples drawn from $p_{lo}^*(\cdot)$. However, we can avoid this sampling step by choosing a loss whose expectation has a closed form. In this paper, we choose $\theta = (c, w) \in \mathbb{R} \times \mathbb{R}^p$ and $\ell(y, x; \theta = (c, w)) = \max(0, 1 - y \cdot (c + w^T x))$. In
other words, given a feature vector \( x \), our model predicts \( y = \text{sign}(c + w^T x) \) and uses the hinge loss with unit margin to measure accuracy. This loss is convex in \( \theta \), 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.**

\[
E_{z \sim p_{lo}^\ast}(y_i = +1, z; \theta) = (1/n_{lo}) \sum_{i : y_i = +1} s_i \cdot \Phi \left( \frac{s_i}{t} \right) + t \cdot \phi \left( \frac{s_i}{t} \right),
\]

\[
s_i = 1 - (c + w^T x_i),
\]

\[
t = \sqrt{w^T \Sigma_{lo} w}.
\]

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 \( s_i \cdot \Phi(s_i/t) + t \cdot \phi(s_i/t) \approx s_i \cdot \mathbb{I}(s_i > 0) = \max(0, s_i) \), 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 \( t \) and \( s \).

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

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

\[
\min_{w \in \mathbb{R}^p, c \in \mathbb{R}} \frac{1}{n_{lo}} \sum_{i : y_i = -1} \max(0, 1 + (c + w^T x_i)) + \frac{1}{n_{lo}} \sum_{i : y_i = +1} \left( s_i \cdot \Phi \left( \frac{s_i}{t} \right) + t \cdot \phi \left( \frac{s_i}{t} \right) \right)
\]

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 \((w_*, c)\) be the minimizer of Eq. 5. Then \( w_* \) lies in the subspace spanned by the \( n_{lo} + n_{hi} \) vectors \( \{x_i\} \).

Hence, when \( n_{lo} + n_{hi} \ll p \) (as in our case), we can assume \( w_* = \sum_{i} \alpha_i x_i \) and only search for the optimal \( \alpha \in \mathbb{R}^{n_{lo} + n_{hi}} \). 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 \( w_* \), 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:

\[
\min_{c' \in \mathbb{R}} \frac{1}{n_{hi}} \sum_{i : y_i = -1} \mathbb{I}(c' + w_{lo}^T x_i) + \frac{1}{n_{lo}} \sum_{i : y_i = +1} E_{z_i \sim \mathcal{N}(x_i, \Sigma_{lo})} \cdot \mathbb{I}(c' + w_{lo}^T z_i) \cdot \Phi \left( \frac{c' + w_{lo}^T x_i}{\sqrt{w_{lo}^T \Sigma_{lo} w_{lo}}} \right) = \Phi \left( \frac{c' + w_{lo}^T x_*}{\sqrt{w_{lo}^T \Sigma_{lo} w_{lo}}} \right)
\]

\[
\begin{align*}
\text{Algorithm 1 DIRECT} & \\
1: & \text{function DIRECT(Training set } X_{lo} \in \mathbb{R}^{n_{lo} \times p}, X_{hi} \in \mathbb{R}^{n_{hi} \times p}) & \\
2: & \Sigma_{lo} \leftarrow \text{robust covariance for minority class (Eq. 2)} & \\
3: & (w_*, c) \leftarrow \text{minimize the convex loss of Eq. 5} & \\
4: & c_* \leftarrow \text{minimize misclassification loss of Eq. 6} & \\
5: & \text{return } (w_*, c_*) & \\
6: & \text{end function}
\end{align*}
\]

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

**Matching desired properties.** DIRECT achieves robustness to limited data (property (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 (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 \( \Sigma_{lo} \) (Eq. 2). We can compute the optimal shrinkage from the data itself, without cross-validation. Thus, we achieve property (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 \( p_{lo}^\ast \) 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 \( x \) (say, a finite number of interaction terms). Then, we can construct new feature vectors \( \bar{x} \) from functions of \( x \), 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
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 5x – 7x 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 3x – 7x 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-
### 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 $n_{lo}$. 

<table>
<thead>
<tr>
<th>$n_{lo}$</th>
<th>$n_{hi}$</th>
<th>Tumors</th>
<th>News Search</th>
<th>Arxiv</th>
<th>MNIST (digits)</th>
<th>MNIST (fashion)</th>
<th>Reuters</th>
<th>20-Newsgroups</th>
<th>Recipes</th>
<th>Tell</th>
<th>UCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>3</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>77.5%*</td>
<td>97.9%*</td>
<td>52.8%*</td>
<td>x</td>
<td>74.5%*</td>
<td>300.7%*</td>
<td>411.0%*</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>474.2%*</td>
<td>474.2%*</td>
<td>474.8%*</td>
<td>78.8%*</td>
<td>101.7%*</td>
<td>51.1%*</td>
<td>33.6%*</td>
<td>77.6%*</td>
<td>474.2%*</td>
<td>623.2%*</td>
</tr>
<tr>
<td>100</td>
<td>7</td>
<td>668.2%*</td>
<td>668.3%*</td>
<td>668.3%*</td>
<td>83.7%*</td>
<td>95.8%*</td>
<td>55.5%*</td>
<td>33.8%*</td>
<td>63.7%*</td>
<td>698.3%*</td>
<td>969.7%*</td>
</tr>
<tr>
<td>1000</td>
<td>5</td>
<td>6.3%</td>
<td>6.3%</td>
<td>6.4%</td>
<td>1.3%</td>
<td>24.4%*</td>
<td>1.8%*</td>
<td>10.9%*</td>
<td>29.4%*</td>
<td>6.4%</td>
<td>29.5%*</td>
</tr>
<tr>
<td>1000</td>
<td>10</td>
<td>7.3%</td>
<td>7.1%</td>
<td>7.3%</td>
<td>2.1%</td>
<td>33.1%*</td>
<td>9.5%*</td>
<td>18.4%*</td>
<td>44.1%*</td>
<td>7.3%</td>
<td>24.7%*</td>
</tr>
<tr>
<td>1000</td>
<td>20</td>
<td>8.1%</td>
<td>7.9%</td>
<td>8.1%</td>
<td>4.2%</td>
<td>37.5%*</td>
<td>9.6%*</td>
<td>16.7%*</td>
<td>55.9%*</td>
<td>8.1%</td>
<td>17.1%*</td>
</tr>
<tr>
<td>1000</td>
<td>50</td>
<td>6.5%</td>
<td>6.3%</td>
<td>6.6%</td>
<td>5.4%</td>
<td>29.8%*</td>
<td>5.9%</td>
<td>15.1%*</td>
<td>73.5%*</td>
<td>6.5%</td>
<td>10.1%*</td>
</tr>
</tbody>
</table>

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


[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.


