

CRCEN: A Generalized Cost-sensitive Neural Network Approach for Imbalanced Classification

Xiangrui Li, Dongxiao Zhu

Department of Computer Science, Wayne State University
{fx7219, dzhu}@wayne.edu

Abstract

Classification on imbalanced datasets is a challenging task in real-world applications. Training conventional classification algorithms directly by minimizing classification error in this scenario can compromise model performance for minority class while optimizing performance for majority class. Traditional approaches to the imbalance problem include re-sampling and cost-sensitive methods. In this paper, we propose a neural network model with novel loss function, CRCEN, for imbalanced classification. Based on the weighted version of cross entropy loss, we provide a theoretical relation for model predicted probability, imbalance ratio and the weighting mechanism. To demonstrate the effectiveness of our proposed model, CRCEN is tested on several benchmark datasets and compared with baseline models.

1 Introduction

Imbalanced datasets are often encountered in real-world problems such computer vision [Buda *et al.*, 2018; Huang *et al.*, 2016], healthcare informatics [Purushotham *et al.*, 2018; Li *et al.*, 2018] and natural language processing [Li *et al.*, 2011; Liu *et al.*, 2006]. In those scenarios, certain classes have abundant training data while some classes only have limited amount of data. The imbalance in training data makes learning classification models a challenging task. Conventional machine learning models treat each training sample equally and be trained by minimizing the overall classification error. Consequently, learned models bias toward correct classification in the majority class over the minority class, resulting in significant performance degradation in the minority class. This phenomenon is not desirable, particularly in domains such as predictions in intensive care units where correct classification of minority samples is critical. Hence, effective learning from imbalanced data is of great importance in machine learning applications.

To tackle the class imbalance problem, different effective strategies have been developed [He and Garcia, 2008]. At the data level, sampling techniques are used as a pre-processing step. Training data are resampled to generate a balanced

dataset, either by oversampling the minority class, or undersampling the majority class (or combine both). Then classification models are trained on the balanced data [Fernandez *et al.*, 2018]. However, sampling methods modify the original data distribution which introduces risks into the following model training: oversampling is at the risk of overfitting while undersampling loses information of majority class. To alleviate the information loss, undersampling can be further combined with ensemble methods where each base classifier is trained on undersampled balanced data.

At the algorithm level, instead of assigning equal weight on all samples, cost-sensitive learning assigns heavier costs on the misclassification of minority samples than majority ones. Classification models are then trained to minimize the total cost. In general, assignment of cost parameters heavily relies on the specific considered problems and there are no general rules for assignment. Common strategies include dynamic cost generation within boosting ensemble methods [Galar *et al.*, 2012], and class-wise reweighting according to class frequency [Aurelio *et al.*, 2019; Castro and Braga, 2013]. In particular, as mentioned in [Castro and Braga, 2013], the class-reweighting cost-sensitive methods are theoretically appealing as the class prior information can be potentially incorporated through the weighting mechanism, consequently improving learning on the minority class.

Recently, solving imbalance problem using (deep) neural networks has attracted much attention. Those works exploit deep neural network’s (DNN) merit of effective feature learning to enhance predictive performances. DNNs are deployed as classifier combining with sampling techniques, cost-sensitive learning or new loss functions [Wang *et al.*, 2016; Chung *et al.*, 2015]. In computer vision where high-level structured information can be learned by convolutional neural networks, properties of learned feature representations are further exploited to improve DNN’s performance [Huang *et al.*, 2016; Dong *et al.*, 2018; Khan *et al.*, 2018].

In this paper, motivated by [Castro and Braga, 2013], we propose a novel cost-sensitive neural networks, Class-wise Reweighted Cross-Entropy Network (CRCEN), to address the imbalanced binary classification problem. In CRCEN, a neural network (in our case, MLP) is used to transform the raw input features into high-level feature representations, based on which final predictions are made using sigmoid

function. Different from the conventional cross entropy loss for binary classification where each training sample is equally weighted, CRCEN imposes different weights on different classes. The reweighting mechanism is capable of promoting CRCEN’s learning on the minority class, enhancing the overall predictive performance. The main contributions of our proposed model are:

- We propose a novel loss function of class-wise reweighted cross entropy based on neural networks to address imbalance classification problem.
- We provide a theoretical derivation on the relation in sample’s predicted probability (once neural network is trained), class weights of loss function and class imbalance ratio. This relation can be generalizable and hold valid for deep neural networks.
- We analyze the generalized relation to gain insights on model performance in imbalanced learning.
- We conduct extensive experiments to demonstrate the effectiveness of our method.

The rest of the paper is organized as follows. In Section 2, we review related works in imbalanced classification problem. Section 3 provide details of the proposed method and analysis on the weighting mechanism. In Section 4, our method is evaluated on several benchmark datasets. Finally, Section 5 concludes the paper with discussion.

2 Related Work

Various approaches on the class imbalance problem have been developed. Here, we focus on two widely-used approaches: sampling methods and cost-sensitive learning. This section also reviews applications of neural networks as our proposed method combines the cost-sensitive learning and MLP. For details of other methods, we refer [He and Garcia, 2008; He and Ma, 2013; Fernandez *et al.*, 2018]

Sampling methods Sampling techniques reduce data imbalance by either oversampling the minority class or undersampling the majority class. [Chawla *et al.*, 2002] proposes the synthetic minority oversampling technique (SMOTE) to oversample by linearly interpolating a pair of close minority samples. ADASYN [He *et al.*, 2008] oversamples according to the local in-class density for each minority sample. To overcome information loss of discarding samples, undersampling is often combined with ensemble methods. [Khoshgof-taar *et al.*, 2007] explores random forest with balanced undersampled dataset in the bagging stage. [Liu *et al.*, 2009] develops balanced bagging and BalanceCascade that dynamically removes easy samples from base classification models.

Cost-sensitive learning As an alternative to sampling methods, cost-sensitive (CS) learning tackle the imbalance problem by imposing heavier costs on misclassified samples. [Tang *et al.*, 2009] develops cost-sensitive SVM with repetitive undersampling to improve the detection of informative samples. Boosting is another popular strategy, due to its internal reweighting strategy in the learning process: boosting dynamically adjusts sample weights for the next iteration according to its previous classification error [Freund and

Schapire, 1997]. In imbalanced learning, models tends to misclassify minority samples at the early stage of boosting. With sample reweighting, minority samples are paid more attention in the later stage and hence performance on minority class can be improved [Sun *et al.*, 2007]. Other variants of boosting are further combined with sampling techniques to improve boosting performance [Galar *et al.*, 2013; Seiffert *et al.*, 2010].

Neural networks in imbalanced learning Neural network is a powerful machine learning model for its merit of high-level feature representation learning. [Dumpala *et al.*, 2018] pairs training samples and uses MLP to predict their labels simultaneously. [Castro and Braga, 2013] develops cost-sensitive MLPs, where loss functions are squared error (L2 loss) and cross entropy respectively, both with each class’s log probability reweighted by inverse class frequency. [Wang *et al.*, 2016] designs new loss functions that encourage equal classification error for majority and minority class for image classification. Additionally, with effective feature extraction of convolutional neural networks, [Huang *et al.*, 2016; Khan *et al.*, 2018; Dong *et al.*, 2018] solve the imbalanced problem in computer vision by exploiting the structured information in high-level features.

Our proposed CRCEN for binary imbalanced classification is a neural network based CS learning method motivated by CS’s theoretical appeal, as mentioned in [Castro and Braga, 2013]. The loss function in CRCEN is the class-wise reweighted cross entropy (CE), and the choice of weight is related to the model’s predictive behavior.

Previous work that is closest to our work in this paper is [Castro and Braga, 2013]. [Castro and Braga, 2013] proposes cost-sensitive MLP (CSMLP) and its loss function is the class-wise weighted squared error loss (i.e. treat classification problem as regression, and label is coded as +1 for majority class and −1 for minority class). Under that specific setting, they derive the same weighting strategy of inverse class frequency to improve learning on the minority class. However, dependence of their weight derivation on the label coding is unclear; it is also difficult to understand the effect if a different weight setting is used. Different from their regression treatment, CRCEN takes the natural probabilistic approach (CE loss) to imbalanced classification problem, which does not rely on a specific label coding. Under some moderate conditions, a relation on MLP’s predicted probability, weight choice and class imbalance ratio is theoretically derived. That relation is further qualitatively analyzed to gain insights on MLP’s predictive performance as well as understand the effect of class weights from a probabilistic perspective. We have noticed that [Aurelio *et al.*, 2019] recently uses weighted CE loss for imbalanced learning, incorporating prior class information heuristically. This is a special case of CRCEN. Furthermore, we provide a theoretical support for this particular weight choice under CRCEN framework.

3 Proposed Method

Notation Let $S = \{(\mathbf{x}_i, y_i) : i = 1, \dots, N\}$ be the imbalanced set of N training samples, where $\mathbf{x}_i \in \mathbf{R}^p$ is the p -dimensional feature vector and $y_i \in \{0, 1\}$ is the class la-

bel. We use $y = 1$ to represent the minority class and $y = 0$ the majority. Further, $S_0 = \{(x_j, y_j) : y_j = 0, j = 1, \dots, N_0\}$ and $S_1 = \{(x_i, y_i) : y_i = 1, i = \dots, N_1\}$ represent the sample sets of majority and minority class respectively with $N_0 \gg N_1$ and $N = N_0 + N_1$. Imbalance ratio (IR) is defined as $N_0/N_1 \gg 1$.

3.1 The CRCEN Loss Function

For the binary classification problem, minimizing cross entropy (CE) (equivalent to maximum likelihood) is an attractive approach due to its modeling of uncertainty, where each training sample is weighted equally. In imbalanced learning, direct use of CE often leads to poor predictive performance. The learning algorithm favors correct classification of samples in the majority class and tends to misclassify the minority samples.

To overcome this limitation and improve correct classification on the minority class, we propose CRCEN to optimize the following class-wise reweighted version of CE loss:

$$\begin{aligned} L(\theta) &= - \sum_{i \in S} \lambda y_i \log p_i(\theta) + (1 - \lambda)(1 - y_i) \log(1 - p_i(\theta)) \\ &= - \lambda \sum_{i \in S_1} \log p_i(\theta) - (1 - \lambda) \sum_{j \in S_0} \log(1 - p_j(\theta)) \\ &= \lambda L_1(\theta) + (1 - \lambda) L_0(\theta), \end{aligned} \quad (1)$$

or the regularized version of $L(\theta)$ (such as L_2 norm of θ):

$$L(\theta) + \beta \Omega(\theta),$$

where $p_i(\theta)$ is the probability $p(y = 1|x_i)$ of x_i belonging to the minority class, modeled by a neural network $f_\theta(x_i)$ with parameter vector θ , $0 < \lambda < 1$ is the weight parameter, β the tuning parameter of Ω . In this paper, f_θ is chosen as MLP.

Note that when $\lambda = 1/2$, CRCEN reduces to the conventional CE loss. Then in the optimization of $L(\theta)$ using gradient descent, the gradient is dominated by the error signal of majority class and decision boundary will be pushed towards minority class. As a consequence, f_θ is likely to have high classification error for minority class. Hence, when greater than $1/2$, λ can be viewed intuitively to strengthen the error signal of minority class in the gradient. Alternatively, optimizing Equation 1 is equivalent to optimizing conventional CE loss, but where training data are rebalanced by simply duplicating each minority sample $\lambda/(1 - \lambda)$ times.

3.2 A Key Equation for Weight λ

Reweight minority samples has been effectively applied in practices. A common strategy is to upweight minority samples by the imbalance ratio IR. In this section, we investigate the theoretical aspect of this weighting mechanism, *with neural network being the predictive models*. As we shall see, the weight λ connects the imbalance ratio with (expected) sample's predicted probability $p(y = 1|x_{\min}) = f_\theta(x_{\min})$ and $p(y = 0|x_{\max}) = 1 - f_\theta(x_{\max})$, where x_{\min} and x_{\max} are minority and majority samples respectively. In the following subsections, we remove θ in analysis for notational brevity.

Assume the output layer of MLP f consists of only one neuron, then the predicted probability is

$$p(y = 1|x) = f(x) = \frac{1}{1 + \exp(-o_x)}, \quad (2)$$

$$o_x = b + \mathbf{W}^T \cdot \mathbf{h}_x,$$

where o is the input to the output neuron, \mathbf{h}_x is the feature embedding of x from the last hidden layer, \mathbf{W} is a parameter subvector of θ for the output neuron and b be the corresponding bias term.

After MLP is trained, the loss function $L(\theta)$ is minimized with optimal solution θ^* . Here, we focus on $L(\theta)$ for simplicity. By optimization theory, we have a necessary condition at $\theta = \theta^*$:

$$\frac{\partial L}{\partial \theta} = \mathbf{0} \iff \lambda \frac{\partial L_1}{\partial \theta} + (1 - \lambda) \frac{\partial L_0}{\partial \theta} = \mathbf{0}.$$

For simplicity, we consider one component w of θ , at $w = w^*$:

$$\begin{aligned} \lambda \frac{\partial L_1}{\partial w} + (1 - \lambda) \frac{\partial L_0}{\partial w} &= 0, \\ \frac{\partial L_1}{\partial w} &= - \sum_{i \in S_1} \frac{1}{p_i} \frac{\partial p_i}{\partial w}, \\ \frac{\partial L_0}{\partial w} &= \sum_{j \in S_0} \frac{1}{1 - p_j} \frac{\partial p_j}{\partial w}, \end{aligned} \quad (3)$$

where $p_i = f(x_i) = p(y = 1|x_i)$ is the predicted probability of sample x_i belonging to minority class.

Since for sigmoid function $\sigma(x) = 1/(1 + \exp(-x))$, its derivative is

$$\sigma'(x) = \sigma(x)(1 - \sigma(x)).$$

Using chain rule, we have the following:

$$\begin{aligned} \frac{\partial p_i}{\partial w} &= \frac{\partial p_i}{\partial o_i} \frac{\partial o_i}{\partial w} = p_i(1 - p_i) \frac{\partial o_i}{\partial w}, \\ \frac{\partial p_j}{\partial w} &= \frac{\partial p_j}{\partial o_j} \frac{\partial o_j}{\partial w} = p_j(1 - p_j) \frac{\partial o_j}{\partial w}. \end{aligned} \quad (4)$$

Plug Equation (4) into Equation (3), we have:

$$-\lambda \sum_{i \in S_1} (1 - p_i) \frac{\partial o_i}{\partial w} + (1 - \lambda) \sum_{j \in S_0} p_j \frac{\partial o_j}{\partial w} = 0. \quad (5)$$

Since Equation (5) holds for any component of parameter θ , we specifically consider the case $w = b$ for the bias term b in Equation (2). Hence $\partial o_i / \partial b = \partial o_j / \partial b = 1$. From Equation (5), we obtain the key equation of CRCEN:

$$\frac{\sum_{i \in S_1} 1 - p_i}{\sum_{j \in S_0} p_j} = \frac{1 - \lambda}{\lambda}, \quad (6)$$

which reveals the relation of weight λ , training sample's predicted probability $p_i (i \in S_1)$ and $p_j (j \in S_0)$ for the minority and majority class, after the neural network is trained.

In the neural network training, L_2 regularization is often applied to prevent overfitting. If the bias term b is not regularized in L_2 term (which is the case usually), Equation (6) still holds. Let $\theta = (\mathbf{W}_h, \mathbf{b})$, where \mathbf{W}_h is the parameter vector for hidden layers and \mathbf{b} the vector of all bias terms. L_2 regularization $\Omega(\theta) = \|\mathbf{W}_h\|_2^2$. Then we have for any component b of \mathbf{b} , $\partial \Omega / \partial b = 0$. Hence Equation (5) holds for $w = b$.

3.3 Theoretical Analysis on Equation (6)

The relation given by Equation (6) depends on the training dataset $S = S_1 \cup S_0$. For theoretical analysis, we make a moderate assumption that, in both the training and testing data, the majority and minority samples are generated from the same class-conditional distribution $P_1(\mathbf{x}|y=1)$ and $P_1(\mathbf{x}|y=0)$ respectively (i.e. no distribution shift between training and testing). Hence, Equation (6) can be generalized as

$$\begin{aligned} \frac{1 - \bar{p}_1}{\bar{p}_0} &\approx \frac{N_0(1 - \lambda)}{N_1\lambda}, \\ \sum_{j \in S_0} p_j &\approx N_0\bar{p}_0, \quad \sum_{i \in S_1} 1 - p_i \approx N_1(1 - \bar{p}_1), \\ \bar{p}_1 &= \mathbf{E}_{\mathbf{x} \sim P_1(\mathbf{x}|y=1)} f_{\theta^*}(\mathbf{x}), \\ \bar{p}_0 &= \mathbf{E}_{\mathbf{x} \sim P_0(\mathbf{x}|y=0)} f_{\theta^*}(\mathbf{x}), \end{aligned} \quad (7)$$

where $P(\mathbf{x}|y=1)$ and $P(\mathbf{x}|y=0)$ are the distributions of the minority and majority class, \mathbf{E} is the expectation operator. Hence, $1 - \bar{p}_1$ is the expected probability of a minority sample with which the trained neural network predicts it in the majority class; \bar{p}_0 be the expected probability of a majority sample being predicted as a minority sample. Here, we use f_{θ^*} to emphasize the dependence of (7) on the trained neural network model. Note that Equation (7) is a general relation regardless of data imbalance.

Predictive performance of the classifier involves a decision making process given p_i . The conventional approach is to set a probability threshold t , such that $y = 1$ if $f_{\theta^*}(\mathbf{x}) > t$, $y = 0$ otherwise. Here, we take $t = 0.5$ for the following analysis to understand model performance. By assuming that training and testing data follow the same distribution, Equation (7) is generalizable from training to testing.

When $\lambda = 1/2$

CRCEN reduces to the conventional cross entropy loss. When imbalance ratio is high in the training data, say $N_0/N_1 = 10$, Equation (7) is $1 - \bar{p}_1 = 10\bar{p}_0$. Since $\bar{p}_1 \geq 0$, we must have

$$\bar{p}_0 \leq 0.1 \iff 1 - \bar{p}_0 \geq 0.9.$$

If $t = 0.5$ is the decision making threshold, this implies that the trained neural network can correctly predict a majority sample, confidently (at least) with probability 0.9, on average.

For prediction on minority class, model performance is more complex. We illustrate the idea with two cases for \bar{p}_0 . Since $\bar{p}_0 \leq 0.1$, again, take $t = 0.5$,

- if $\bar{p}_0 = 0.08$, then we have $\bar{p}_1 = 0.2$. That implies the predicted probability of a minority sample being minority is 0.2 on average. Hence, the classifier must misclassify most minority samples ($0.2 < 0.5$), resulting in very poor predictive accuracy for minority class.
- if $\bar{p}_0 = 0.04$, then $\bar{p}_1 = 0.6$. Hence, the classifier can achieve good performance for minority samples. But the extent of “goodness” depends on the trained network f_{θ^*} and $P_0(\mathbf{x}|y=1)$, i.e. the variance $\mathbf{V}_{\mathbf{x} \sim P_1(\mathbf{x}|y=1)}(f_{\theta^*}(\mathbf{x}))$. If the variance is small (relative to the average distance from the decision boundary in

λ	$\frac{1}{2}$	$\frac{N_0}{N_0+N_1}$	$\frac{2N_0}{(2N_0+N_1)}$
RHS	10	1	0.5
LHS (Sim1)	10.05 (1.13)	1.00 (0.09)	0.50 (0.04)
LHS (Sim2)	10.12 (0.67)	1.01 (0.05)	0.50 (0.03)

Table 1: Simulation results (along with standard deviation) for Equation (7) over 100 runs. RHS represents theoretical value on the right hand side of (7); LHS the simulated value on the left hand side.

the latent feature space), classifier can still achieve very high accuracy. If large, performance would degrade.

Geometrically, the imbalance of training data would push decision boundary toward to minority class in the latent feature space learned by neural network.

The analysis above theoretically explains why classifier always has good performance for majority class, as well as how performance for minority class is connected with model training and data distribution in imbalanced learning.

When $\lambda = N_0/(N_0 + N_1)$

This strategy is equivalent to the empirical weighting by inverse class frequency deployed in [Aurelio *et al.*, 2019; Castro and Braga, 2013]. With this choice of λ , Equation (7) is $\bar{p}_1 = 1 - \bar{p}_0$. That is, the probability of predicting a minority sample as minority is equal to the probability of predicting a majority sample as majority. Taking $t = 0.5$, good predictive performance for both minority and majority class is guaranteed. However, the extent of “goodness” depends on $\mathbf{V}_{\mathbf{x} \sim P_1(\mathbf{x}|y=1)}(f_{\theta^*}(\mathbf{x}))$ and $\mathbf{V}_{\mathbf{x} \sim P_0(\mathbf{x}|y=0)}(f_{\theta^*}(\mathbf{x}))$.

For General Choice of λ

λ is the parameter controlling the ratio between probabilities $1 - \bar{p}_1$ and \bar{p}_0 . For example, when $\lambda = 2N_0/(N_1 + 2N_0)$, $1 - \bar{p}_1 = \bar{p}_0/2$. The loss function in CRCEN then deploys a weight setting equivalent to:

$$2N_0L_1(\theta) + N_1L_0(\theta).$$

Since $0 \leq \bar{p}_0 \leq 1$, $\bar{p}_1 \geq 0.5$. We are guaranteed to have good predictive performance for minority class (when $t = 0.5$). Assume $\bar{p}_0 = 0.4$ ¹ (i.e. $1 - \bar{p}_0 = 0.6$), we obtain $\bar{p}_1 = 0.8$, which implies the prediction accuracy for minority class can be possibly boosted at the cost of a small accuracy degradation, but still maintaining good performance, for majority class. For the exact relation, we plan to investigate this problem in our future studies.

3.4 Simulations for Correctness of Equation (7)

In order to check the correctness of Equation (7), we conduct simulations under two settings. The imbalance ratio is 10 in training data ($N_1 = 1000$, $N_0 = 10000$), testing data size is (1000, 1000); both training and testing data follow the same data distribution.

¹The true value of \bar{p}_0 is generally unknown as it depends on the true data distribution.

Dataset	# Sample	# Feature	IR
Abalone	4177	10	9.7
Coil	9822	85	16
Satimage	6435	36	9.3
Scene	2407	294	13
Solar	1389	32	19
USCrime	1994	100	12

Table 2: Details of datasets. IR represents imbalance ratio between majority class and minority class.

- Sim1: $P_1(x|y = 1) = \mathcal{N}(-1.5, 1) + \mathcal{U}(0, 0.5)$, $P_2(x|y = 0) = \mathcal{N}(1.5, 1) + \mathcal{U}(-0.5, 0)$. Logistic regression is fitted (as a special case of MLP).
- Sim2: $P_1(x|y = 1) = \mathcal{N}(\mu_1, \sigma_1)$, $P_0(x|y = 0) = \mathcal{N}(\mu_0, \sigma_0)$, where $\mu_1 = (0, 0, 0)$, $\mu_0 = (1, 1, 1)$, $\sigma_1 = 1.2I$, $\sigma_0 = I$. A one-hidden-layer MLP of layer size (3, 10, 1) and sigmoid activation is fitted.

λ is the weight used in CRCEN and the latter is then trained on the training data. The predicted probability for testing data is calculated and used in the LHS of Equation (7) to approximate \bar{p}_1 and \bar{p}_0 . Table 1 shows simulation results under three λ settings. We see from the table that simulated values match with the theoretical values accurately, demonstrating the correctness of Equation (7).

4 Experiments

In this section, we evaluate CRCEN on real-world datasets with various imbalance degrees that are widely used in imbalanced learning. All datasets tested in the experiments are extracted from the “imbalanced-learn” Python package [Lemaître *et al.*, 2017]. Table 2 shows the details of the datasets used in the experiments.

For performance comparison, we test several baseline models for imbalanced learning, including sampling methods SMOTE [Chawla *et al.*, 2002] and ADASYN [He *et al.*, 2008], ensemble methods Rusboost (RUSB) [Seiffert *et al.*, 2010], balanced random forest (BRF) [Khoshgoftaar *et al.*, 2007] and EasyEnsemble (EE) [Liu *et al.*, 2009], and neural network-based cost-sensitive method CSMLP [Castro and Braga, 2013]. As sampling methods are a data preprocessing step, we use MLP as the classifier after training data are re-sampled. For ensemble methods, default base classifiers are used. In addition, a MLP classifier trained on the original data with conventional cross entropy loss is tested. Most methods are implemented in imbalanced-learn and sklearn package. For our proposed CRCEN, we use MLP as the classifier and implement it in Pytorch. All MLPs are fixed with 3 layers with number of hidden neurons selected for each dataset.

In the experiments, each dataset is divided into training and testing sets by a stratified split 0.75/0.25 to ensure imbalance ratio is maintained. All models are trained on the training data and model performance is evaluated on the testing data. We repeats train/test split 4 times. We select model parameters (including L_2 regularization, number of hidden neurons) using 4-fold cross validation with a grid search on the training data in the first run of the experiment and then fix them

in the subsequent runs. This procedure is used to test model robustness to the variations in train/test splits.

As the overall accuracy is known misleading for imbalanced datasets, we use F-measure (F1) and Gm as evaluation metrics. F-measure and G-mean(Gm) are defined as follows:

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$

$$Gm = \sqrt{\text{Recall} \cdot \text{Specificity}},$$

where Precision = TP/(TP+FP), Recall = TP/(TP+FN) and Specificity = TN/(FP+TN), TP represents true positive, FP false positive, FN false negative and TN true negative.

4.1 Predictive Results

Predictive performance on the testing data is reported in Table 3. As we can see from the table, CRCEN has an overall better performance than other methods. Compared with MLP trained directly on the training data, all methods achieve great improvements in G-mean. Since MLP tends to misclassify minority samples into majority class, this results in a large error and low recall score for minority class. Consequently, the G-mean of MLP is low. On the contrary, this demonstrates that all those techniques can effectively strengthen classifier’s detection for minority class and improve overall performance. CRCEN and CSMLP are both cost-sensitive methods based on MLP with the same weighting mechanism. We see that in the experiments CRCEN has slightly better but comparable performances with CSMLP. As explained in previous sections, both methods increase learning of minority samples with theoretical guarantees, by learning a balanced boundary between two classes. However, CRCEN’s probabilistic approach is more appropriate and effective for classification problems [Lee *et al.*, 2015]. Rusboost (RUSB) is an ensemble cost-sensitive method that costs are dynamically assigned to misclassified samples. From the table, RUSB’s performance is not as good as CRCEN and CSMLP. By checking its prediction (results not shown here), RUSB achieves highest classification accuracy for minority class, at the cost of significant amount of misclassification in majority class. This is because samples on the boundary can be easily misclassified, RUSB would assign a large cost to the boundary samples. Consequently, decision boundary will be pushed towards majority class and results in decreasing in precision and specificity, hence degradation in overall performance (F1 and G-mean).

4.2 Effect of λ

In many real-world applications, correct classification of minority samples ($y = 1$) is of high priority over misclassification of majority samples ($y = 0$), making imbalanced learning extremely useful. In terms of evaluation metrics, a predictive model that has high recall is preferred. For cost-sensitive learning, imposing more cost on minority samples would generally improve recall. However, since there is tradeoff between recall and precision (or specificity), higher cost will decrease precision and (or specificity). Under CRCEN, the weight parameter λ controls the tradeoff. In this section, we investigate this relation.

	Abalone		Coil		Satimage		Scene		Solar		UScrime	
	F1	Gm	F1	Gm	F1	Gm	F1	Gm	F1	Gm	F1	Gm
MLP	0	0	0.107	0.273	0.554	0.668	0.211	0.341	0	0	0.475	0.597
CSMLP	0.392	0.798	0.200	0.658	0.611	0.890	0.259	0.662	0.240	0.655	0.472	0.707
CRCEN	0.400	0.808	0.205	0.678	0.626	0.894	0.298	0.701	0.264	0.690	0.525	0.726
ADASYN	0.389	0.804	0.166	0.411	0.583	0.898	0.253	0.512	0.240	0.643	0.448	0.652
SMOTE	0.400	0.808	0.161	0.406	0.624	0.891	0.234	0.487	0.232	0.647	0.466	0.675
RUSB	0.385	0.792	0.163	0.624	0.518	0.874	0.163	0.561	0.174	0.705	0.407	0.831
EE	0.377	0.789	0.200	0.670	0.537	0.869	0.256	0.707	0.169	0.670	0.434	0.858
BRF	0.393	0.794	0.196	0.674	0.579	0.890	0.256	0.709	0.180	0.699	0.459	0.865

Table 3: Predictive performance on the testing data. F1 is F-measure and Gm is G-mean.

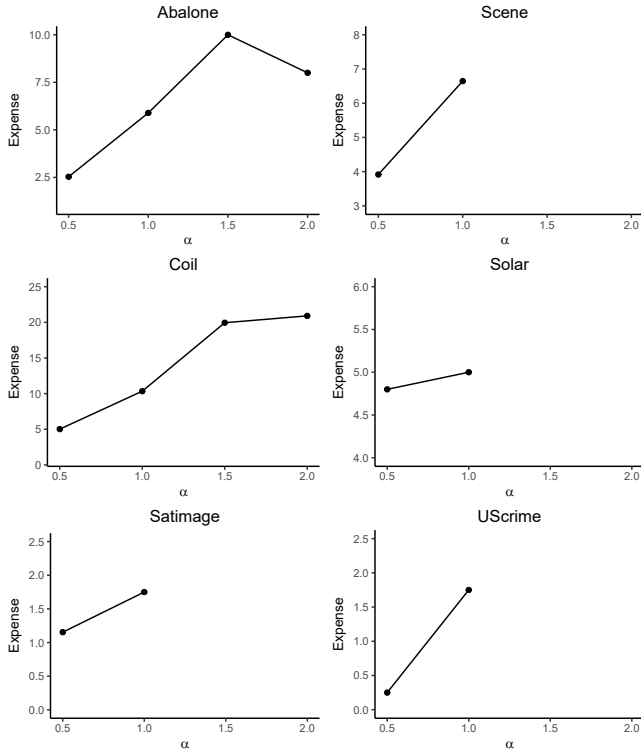


Figure 1: Expense plot for different λ settings corresponding to α s. Expense measures the tradeoff between FNs and FPs when heavier cost is added to minority class in CRCEN. When $\alpha = 0.5$, plot corresponds to the expense from conventional cross entropy loss ($\lambda = 1/2$) to cost sensitive loss CRCEN.

We set $\lambda = \frac{1}{2}$ and $\frac{\alpha N_0}{(\alpha N_0 + N_1)}$, $\alpha = (0.5, 1, 1.5, 2)$ to see to what extent of increase in weight can lead to performance gain on classifying the minority class, and how much performance loss on magnitude of the majority class. Note that when $\lambda = 1/2$, CRCEN is equivalent to conventional cross entropy loss. For each value of α , we train a MLP for classification. Once the models are trained, predictions are made on testing data. To quantify the tradeoff between re-

call and precision, we define *expense* as the ratio of change of FPs to that of FNs between two consecutive α s (since α increases, number of FN decreases and FP increases). For example, (32, 158) and (13, 263) are two pairs of (FN, FP) corresponding to $\alpha = 0.5$ and $\alpha = 1$ respectively, predicted on the testing data of Abalone. Then the budget is $(263 - 158)/(32 - 13) = 5.53$. This ratio can be viewed as the budget that one can afford for detecting a false negative sample. The decision making threshold is $t = 0.5$.

Figure 1 shows the expense plots for six datasets. From the table, we can observe a trend that the expense increases as α (namely λ) increases. This implies that by imposing heavier costs, we can improve detection of true positives however at an increasing cost of false positives. When $\alpha > 1$, in Scene, Solar, Satimage and UScrime datasets, heavier costs don't improve classifier's performance on minority class. With limited amount of training data, this in turn increases the risk of overfitting, resulting in performance degradation. We see that for datasets Satimage and UScrime, the expense is relatively small (less than 2 when $\alpha = 1$), compared with Abalone, Scene, Coil and Solar. In Table 3, MLP already has good performance in those two datasets. With a low expense and high class imbalance ratio, cost-sensitive can further improve the model overall performance, as confirmed in Table 3. From the same perspective, CRCEN is also effective for more complex datasets Abalone, scene, Coil and Solar.

5 Conclusion and Discussion

In this paper, we proposed a novel neural-network based model CRCEN for imbalanced learning problem. The object function in CRCEN is a class-wise reweighed version of the cross entropy loss. With this simple form, under some mild conditions, we derive a non-trivial probabilistic relation that can help us understand model's predictive behavior. When the weights are set to inverse class frequency as a heuristic, the derived relation provides explanation for the effectiveness of this approach with theoretical guarantees. Extensive experiments are conducted to demonstrate the effectiveness of CRCEN. For future studies, we plan to investigate the relation for a general choice of λ to understand how model performance is affected accordingly.

References

- [Aurelio et al., 2019] Yuri Sousa Aurelio, Gustavo Matheus de Almeida, Cristiano Leite de Castro, and Antonio Padua Braga. Learning from imbalanced data sets with weighted cross-entropy function. *Neural Processing Letters*, pages 1–13, 2019.
- [Buda et al., 2018] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.
- [Castro and Braga, 2013] Cristiano L Castro and Antônio P Braga. Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data. *IEEE transactions on neural networks and learning systems*, 24(6):888–899, 2013.
- [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, 2002.
- [Chung et al., 2015] Yu-An Chung, Hsuan-Tien Lin, and Shao-Wen Yang. Cost-aware pre-training for multiclass cost-sensitive deep learning. *arXiv preprint arXiv:1511.09337*, 2015.
- [Dong et al., 2018] Qi Dong, Shaogang Gong, and Xiatian Zhu. Imbalanced deep learning by minority class incremental rectification. *IEEE transactions on pattern analysis and machine intelligence*, 2018.
- [Dumpala et al., 2018] Sri Harsha Dumpala, Rupayan Chakraborty, Sunil Kumar Kopparapu, and TCS Reseach. A novel data representation for effective learning in class imbalanced scenarios. In *IJCAI*, pages 2100–2106, 2018.
- [Fernandez et al., 2018] Alberto Fernandez, Salvador Garcia, Mikel Galar, Ronaldo Prati, Bartosz Krawczyk, and Francisso Herrera. *Learning from imbalanced data sets*. Springer, 2018.
- [Freund and Schapire, 1997] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.
- [Galar et al., 2012] Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484, 2012.
- [Galar et al., 2013] Mikel Galar, Alberto Fernández, Edurne Barrenechea, and Francisco Herrera. Eusboost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. *Pattern Recognition*, 46(12):3460–3471, 2013.
- [He and Garcia, 2008] Haibo He and Eduardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, (9):1263–1284, 2008.
- [He and Ma, 2013] Haibo He and Yunqian Ma. *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons, 2013.
- [He et al., 2008] Haibo He, Yang Bai, Eduardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328. IEEE, 2008.
- [Huang et al., 2016] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016.
- [Khan et al., 2018] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8):3573–3587, 2018.
- [Khoshgoftaar et al., 2007] Taghi M Khoshgoftaar, Moiz Golawala, and Jason Van Hulse. An empirical study of learning from imbalanced data using random forest. In *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, volume 2, pages 310–317. IEEE, 2007.
- [Lee et al., 2015] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial Intelligence and Statistics*, pages 562–570, 2015.
- [Lemaitre et al., 2017] Guillaume Lemaitre, Fernando Nogueira, and Christos K Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, 18(1):559–563, 2017.
- [Li et al., 2011] Shoushan Li, Zhongqing Wang, Guodong Zhou, and Sophia Yat Mei Lee. Semi-supervised learning for imbalanced sentiment classification. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.
- [Li et al., 2018] Xiangrui Li, Dongxiao Zhu, and Ming Dong. Multinomial classification with class-conditional overlapping sparse feature groups. *Pattern Recognition Letters*, 101:37–43, 2018.
- [Liu et al., 2006] Yang Liu, Nitesh V Chawla, Mary P Harper, Elizabeth Shriberg, and Andreas Stolcke. A study in machine learning from imbalanced data for sentence boundary detection in speech. *Computer Speech & Language*, 20(4):468–494, 2006.
- [Liu et al., 2009] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2009.
- [Purushotham et al., 2018] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of biomedical informatics*, 83:112–134, 2018.
- [Seiffert et al., 2010] Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. Rusboost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40(1):185–197, 2010.
- [Sun et al., 2007] Yanmin Sun, Mohamed S Kamel, Andrew KC Wong, and Yang Wang. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 40(12):3358–3378, 2007.
- [Tang et al., 2009] Yuchun Tang, Yan-Qing Zhang, Nitesh V Chawla, and Sven Krasser. Svms modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):281–288, 2009.
- [Wang et al., 2016] Shoujin Wang, Wei Liu, Jia Wu, Longbing Cao, Qinxue Meng, and Paul J Kennedy. Training deep neural networks on imbalanced data sets. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 4368–4374. IEEE, 2016.