

# Optimal Rejection Function Meets Character Recognition Tasks

Xiaotong Ji<sup>1</sup>, Yuchen Zheng<sup>1</sup>[0000–0003–3093–6929],  
Daiki Suehiro<sup>1,2</sup>[0000–0001–8901–9063], and Seiichi Uchida<sup>1</sup>[0000–0001–8592–7566]

<sup>1</sup> Kyushu University, 744 Motooka Nishi-ku, Fukuoka 819-0395, Japan  
{xiaotong.ji, yuchen}@human.ait.kyushu-u.ac.jp  
{suehiro, uchida}@ait.kyushu-u.ac.jp  
<sup>2</sup> RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan

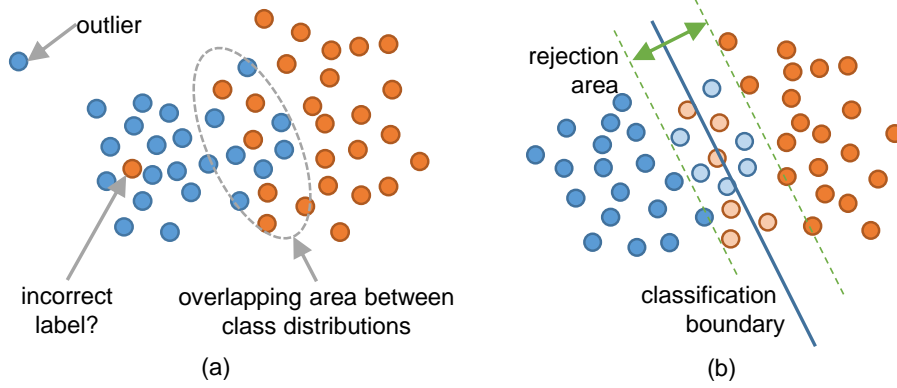
**Abstract.** In this paper, we propose an optimal rejection method for rejecting ambiguous samples by a rejection function. This rejection function is trained together with a classification function under the framework of Learning-with-Rejection (LwR). The highlights of LwR are: (1) the rejection strategy is not heuristic but has a strong background from a machine learning theory, and (2) the rejection function can be trained on an arbitrary feature space which is different from the feature space for classification. The latter suggests we can choose a feature space which is more suitable for rejection. Although the past research on LwR focused only its theoretical aspect, we propose to utilize LwR for practical pattern classification tasks. Moreover, we propose to use features from different CNN layers for classification and rejection. Our extensive experiments of notMNIST classification and character/non-character classification demonstrate that the proposed method achieves better performance than traditional rejection strategies.

**Keywords:** Learning with Rejection · Optimal Rejection Function · Theoretical Machine Learning

## 1 Introduction

According to general performance improvement in various recognition tasks [1,2,3], many users may expect that most test samples should be classified correctly. In fact, recent classifiers use forced decision-making strategies, where any test sample is always classified into one of the target classes [4,5]. However, even with the recent classifiers, it is theoretically impossible to classify test samples correctly in several cases. For example, we cannot guarantee the correct recognition of a sample from the overlapping region of class distributions.

As a practical remedy to deal with those ambiguous samples, recognition with a rejection option has been used so far [6,7]. As shown in Figure 1 (a), there are three main targets for rejection: outliers, samples with incorrect labels, and samples in an overlapping area of class distributions. Among them, outliers and incorrect labels are detected by anomaly detection methods. The overlapping



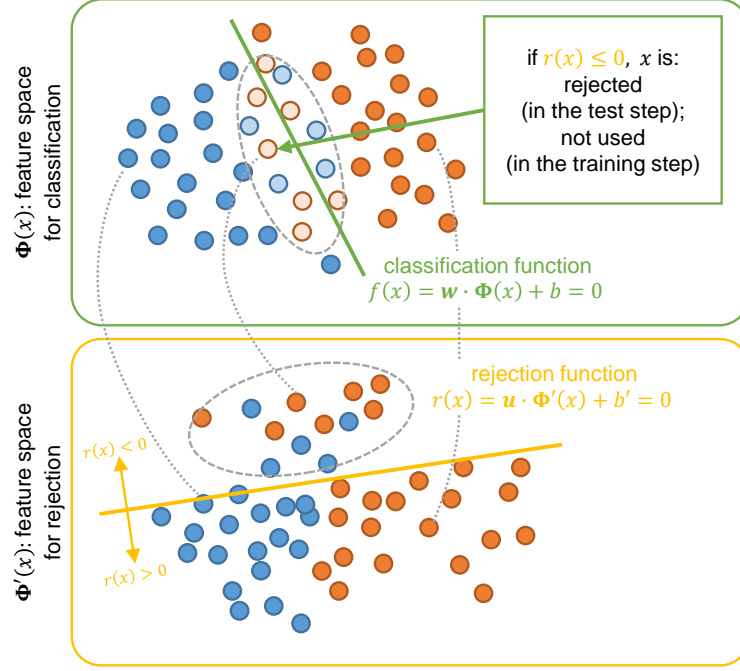
**Fig. 1.** (a) Three main targets for rejection. (b) Typical heuristics for rejecting the samples in the overlapping area of class distributions. This method assumes that we already have an accurate classification boundary — but, how can we reject ambiguous samples when we train the classification boundary?

area is a more common reason to introduce the rejection option into the training stage as well as the testing stage, since the samples in the overlapping area are inherently distinguishable and thus better to be treated as “don’t-care” samples.

In the long history of pattern recognition, researchers have tried to reject ambiguous samples in various ways. The most typical method is to reject samples around the classification boundary. Figure 1 (b) shows a naive rejection method for a support vector machine (SVM) for rejecting samples in the overlapping area. If the distance between a sample and the classification boundary is smaller than a threshold, the sample is going to be rejected [8].

Although this simple rejection method is often employed for practical applications, it has a clear problem that it needs an accurate classification boundary in advance to rejection. In other words, it can reject only test samples. Ambiguous samples in the overlapping area also exist in the training set and they will badly affect the classification boundary. This fact implies that we need to introduce a more sophisticated rejection method where the rejection criterion should be “co-optimized” with the classification boundary in the training stage.

In this paper, we propose an optimal rejection method for rejecting ambiguous training and test samples for tough character recognition tasks. In the proposed method, a rejection function is optimized (i.e., trained) together with a classification function during the training step. To realize this co-optimization, we follow the *learning-with-rejection* (LwR) framework proposed in theoretical machine learning research [9]. Figure 2 illustrates LwR framework. During the training step, several samples are determined to be rejected by a rejection function and they are not used to train the classifier function. It should be emphasized again this is a co-optimization framework where the rejection function and the classification function are optimized by solving a single optimization problem. LwR has many promising characteristics; for example, it has a strong theoret-



**Fig. 2.** An overview of *learning-with-rejection* for a two-class problem. The rejection function  $r(x)$  is co-trained with the classification function  $f(x)$ . Note that different feature spaces are used for optimizing  $r(x)$  and  $f(x)$ .

ical support about its optimality (about the overfitting risk) and it can use a different feature space for its rejection function.

The main contributions of our work are summarized as follows:

- This is the first application of LwR to a practical pattern recognition task. Although the idea of LwR is very suitable to various pattern recognition tasks (including, of course, character recognition and document processing), no application research has been done so far, to the authors' best knowledge.
- For designing features for classification and rejection, we utilize a convolutional neural network (CNN) as a multiple-feature extractor; the classification function is trained by using the output from the final convolution layer and the rejection function is trained from a different layer.
- We conducted two classification experiments to show that LwR outperforms CNN with confidence-based rejection, and traditional SVM with the rejection strategy of Figure 1 (b).

The rest of the paper is organized as follows: Section II reviews related works and Section III is devoted to explain the problem formulation and the theoretical background of LwR. In Section IV, the experimental setting and the discussion of final results are shown, while conclusion is given in Section V.

## 2 Related work

Since data usually contains unrecognizable samples, data cleansing [10,11,12] has been an important topic in many fields such as medical image processing [13,14], document analysis [15,16,17] and commercial applications [18]. Among those applications, ambiguous samples in dataset might incur serious troubles at data processing stage. To handle this problem, many data cleansing methods have been studied [19,20] and the goal of all of them is to keep the data clean. To guarantee the data integrity, most of them apply repairing operation to the target data with inconsistent format. Instead of repairing those acceptable flaws, authors in [21] aim to remove erroneous samples from entire data, and in [22], Fecker et al. propose a method to reject query manuscripts without corresponding writer with a rejection option. These works show the necessity of rejection operation in data cleansing. For machine learning tasks, training data with undesirable samples (as shown in Figure 1 (a)) could lead to a ill-trained model, whose performance in the test phase is doomed to be poor. On the other hand, test data could also contains such unwelcome samples but are apparently better to be removed by the trained model than those data cleansing methods.

Other than data cleansing, rejection operation is also commonly used in many fields of machine learning such as handwriting recognition [23,24], medical image segmentation [25,26] and image classification tasks [27,28,29]. Here, the rejection operation are done during the test stage but not the training stage. In the handwriting recognition field, Kessentini et al. apply a rejection option to Latin and Arabic scripts recognition and gained significant improvements compared to other state-of-the-art techniques [30]. Mesquita et al. develop a classification method with rejection option [31] to postpone the final decision of non-classified modules for software defect prediction. In [32], He et al. construct a model with rejection on handwritten number recognition, where the rejection option nicely prevented misclassifications and showed high reliability. Mekhmoukh and Mokrani employ a segmentation algorithm that is very sensitive to outliers in [26] to effectively reject anomalous samples. Niu et al. in [33] present a hybrid model of integrating the synergy of a CNN and SVM to recognize different types of patterns in MNIST, achieved 100% reliability with a rejection rate of 0.56%. However, works mentioned above treat rejection operation as the extension of classification function. It is notable that dependence on classification function could endow potential limitation to rejection ability of the models.

Rejection operation also plays an important role in document processing field [34,35]. Bertolami et al. investigate rejection strategies based on various confidence measures in [36] for unconstrained offline handwritten text line recognition. In [37], Li et al. propose a new CNN based confidence estimation framework to reject the misrecognized character whose confidence value exceeds a preset threshold. However, to our knowledge, there still has no work on combining CNNs and independent rejection functions in document analysis and recognition field. Since it is well known that CNNs have powerful feature extraction capabilities, outputs of hidden layers in CNNs could be used instead of kernels as feature spaces for model training. In this paper, we propose a novel frame-

work with a rejection function learned along with but not dependent on the classification function, using features extracted from a pre-trained CNN.

### 3 Learning with Rejection

In this section, we introduce the problem formulation of LwR. Let  $(x_1, y_1), \dots, (x_m, y_m)$  denote training samples. Let  $\Phi$  and  $\Phi'$  denote functions that map  $x$  to different feature spaces. Let  $f = \mathbf{w} \cdot \Phi(x) + b$  be a linear classification function over  $\Phi$ , while  $r = \mathbf{u} \cdot \Phi'(x) + b'$  be a linear rejection function over  $\Phi'$ . Our LwR problem is formulated as an optimization problem of  $\mathbf{w}$ ,  $b$ ,  $\mathbf{u}$  and  $b'$  as follows. As shown in Figure 2, a two-class classification problem with reject option uses the following decision rule :

$$\begin{cases} +1 \text{ (} x \text{ is class1)} & \text{if } (f(x) > 0) \wedge (r(x) > 0), \\ -1 \text{ (} x \text{ is class2)} & \text{if } (f(x) \leq 0) \wedge (r(x) > 0), \\ \text{reject} & \text{if } r(x) \leq 0. \end{cases}$$

To achieve a good balance between the classification accuracy and the rejection rate, we minimize the following risk:

$$R(f, r) = \sum_{i=1}^m (1_{(\text{sgn}(f(x_i)) \neq y_i) \wedge (r(x_i) > 0)} + c 1_{(r(x_i) \leq 0)}), \quad (1)$$

where  $c$  is a parameter which weights the rejection cost. The cost represents the penalty to the rejection operation.

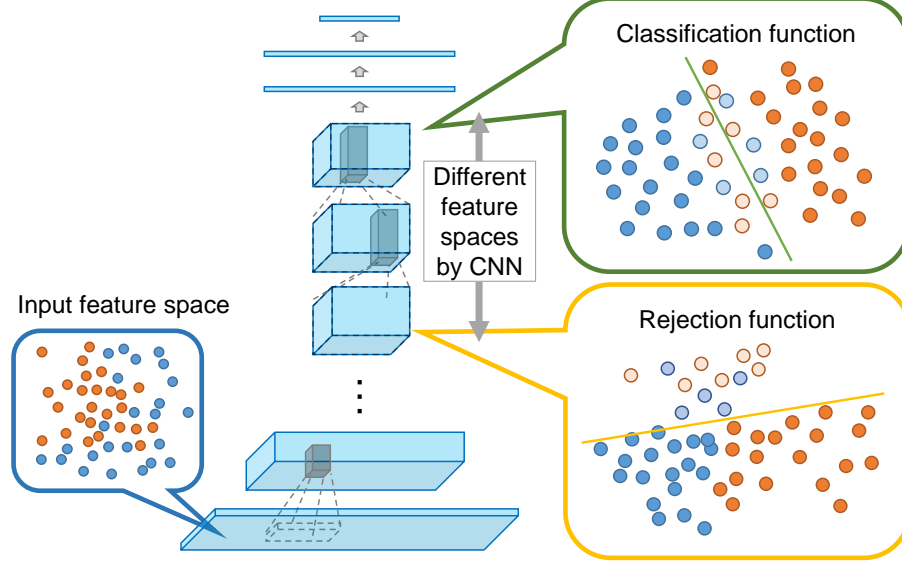
As an example in document processing task, assuming digit recognition for bank bills. In this case, there is almost no tolerance to errors and thus the rejection cost  $c$  must be set at a lower value for rejecting more suspicious patterns to avoid possible mistakes. In contrast, in the case of character recognition of personal diary,  $c$  can be set at a larger value for accepting more ambiguous characters is acceptable.

Since the direct minimization of the risk  $R$  is difficult, Cortes et al. [9] proposed a convex surrogate loss  $L$  as follows:

$$L = \max \left( 1 + \frac{\alpha}{2}(r(x) - yf(x)), c(1 - \beta r(x)), 0 \right),$$

where  $\alpha = 1$  and  $\beta = 1/(1 - 2c)$ . In [9], it is proved that the minimization problem of  $L$  along with a typical regularization of  $\mathbf{w}$  and  $\mathbf{u}$  results in the following SVM-like optimization problem:

$$\begin{aligned} \min_{\mathbf{w}, \mathbf{u}, b, b', \xi} \quad & \frac{\lambda}{2} \|\mathbf{w}\|^2 + \frac{\lambda'}{2} \|\mathbf{u}\|^2 + \sum_{i=1}^m \xi_i \\ \text{sub. to} \quad & \xi_i \geq c(1 - \beta(\mathbf{u} \cdot \Phi'(x_i) + b')), \\ & \xi_i \geq 1 + \frac{\alpha}{2}(\mathbf{u} \cdot \Phi'(x_i) + b' - y_i(\mathbf{w} \cdot \Phi(x_i) + b)), \\ & \xi_i \geq 0 \quad (i = 1, \dots, m), \end{aligned} \quad (2)$$



**Fig. 3.** Extracting different features from CNN for LwR.

where  $\lambda$  and  $\lambda'$  are regularization parameters and  $\xi_i$  is a slack variable. This is a minimization problem of a quadratic objective function with linear constraints and thus its optimal solution can be easily obtained by using a quadratic programming (QP) solver.

It should be emphasized that LwR solution of the above-mentioned formulation has a strong theoretical background. Similarly to the fact that the standard SVM solution has theoretical guarantee of its generalization performance, the optimal solution of the optimization problem (2) in LwR mentioned above also has a theoretical support. See Appendix for its brief explanation. It should also be underlined that the original LwR research has mainly focused on this theoretical analysis and thus has never been used for any practical application. To the authors' best knowledge, this paper is the first application of LwR to a practical task.

## 4 Experimental result

In this section, we demonstrate that our optimal rejection approach effectively works for character recognition tasks. More precisely, we consider two binary classification tasks: classification of similar characters and classification of character and non-character. Of course, in those tasks, it is better to achieve a high accuracy with a low reject rate.

#### 4.1 Experimental Setup

**Methods for comparison** Through the experiments, we compare three methods: CNN classifier, SVM classifier<sup>3</sup>, and LwR approach. The SVM classifier is set linear for deep CNN features we use already contain nonlinear compositions.

The performances of these three classification models are compared with shifting rejection cost parameters as  $c = 0.1, \dots, 0.4$ . CNN and SVM output class probability  $p(y|x)$ , and thus we consider their final output  $f(x)$  with reject option based on confidence threshold  $\theta$ , that is,

$$f(x) = \begin{cases} +1 \text{ (} x \text{ is class1)} & \text{if } p(+1|x) > \theta, \\ -1 \text{ (} x \text{ is class2)} & \text{if } p(-1|x) > \theta, \\ \text{reject} & \text{otherwise.} \end{cases}$$

We determine the optimal confidence threshold  $\theta$  of rejection according to the risk  $R$  using the *trained*  $f(x)$  and a validation set, where the  $R$  of CNN and SVM can be defined as:

$$R(f) = \sum_{i=1}^m (1_{((f(x_i)) \neq y_i) \wedge (f(x) \neq \text{reject})} + c 1_{(f(x) = \text{reject})}).$$

For the LwR model, validation samples are used to tune hyperparameters  $\lambda$  and  $\lambda'$  for LwR could determine its rejection rule by using only training samples.

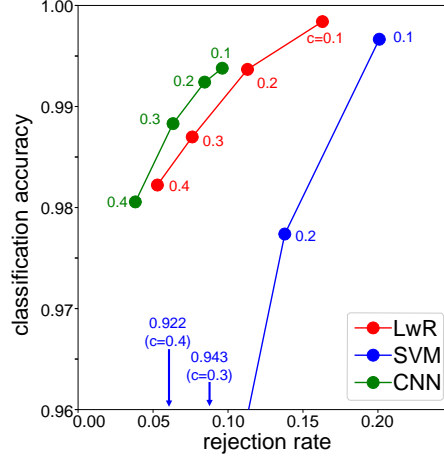
The learning scenario of these three methods in these experiments is introduced as follow. For each binary classification tasks, we divide the binary labeled training samples into three sets: (1) Sample set for training a CNN. This CNN is used as not only the feature extractor for LwR and SVM but also a CNN classifier. (2) Sample set for training SVM and LwR with the features extracted by the trained CNN. (3) Sample set for validating the parameters of each method.

**Feature Extraction** As mentioned above, one of the key ideas of LwR is to use different feature spaces to construct rejection function. For this point, we use the outputs of two different layers of the trained CNN for the classifier  $f(x)$  and rejection function  $r(x)$  respectively, as shown in Figure 3. Specifically, we use the final convolutional layer of the CNN for classification. Note that the SVM classifier also uses the same layer as its feature extractor. For the training stage of rejection function, the second final layer of the CNN is used for we found it achieved good performance in our preliminary experiments. In addition, if the same layer is used for both classification and rejection functions, the LwR model will degrade to a SVM-like model with information from only one feature space.

**Evaluation criteria** Using sample set  $S$  (i.e. the test set) for each task, we evaluate the classification performance with reject option of these three models by following metrics: classification accuracy for non-rejected test samples (which we simply call *accuracy* for short) and *rejection rate* are defined as:



**Fig. 4.** Examples of “I” (upper row) and “J” (lower row) from notMNIST.



**Fig. 5.** The relationship between accuracy and reject rate for the classification task between “I” and “J” from notMNIST dataset.

$$\text{accuracy} = \frac{\sum_{x_i \in \{S - S_{\text{rej}}\}} 1_{f(x_i)=y_i}}{|S| - |S_{\text{rej}}|}, \quad \text{rejection rate} = \frac{|S_{\text{rej}}|}{|S|},$$

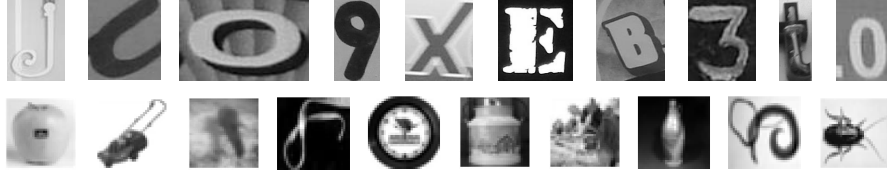
where  $S_{\text{rej}}$  is the set of rejected samples, and  $f(x_i)$  is the classification function of each method.

## 4.2 Classification of Similar Characters with Reject Option

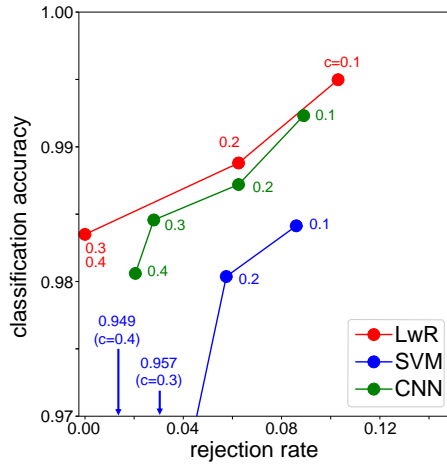
In this section, we use a dataset named notMNIST<sup>4</sup> as both training and test data. The notMNIST dataset is an image dataset consists of A-to-J alphabets including various kinds of fonts, whose examples are shown in Figure 4. We choose a pair of very similar letters “I” and “J” to organize a tough two-class recognition task, where reject operation of many ambiguous samples is necessary. We randomly selected 3,000 training samples of “I” and “J” respectively from

<sup>3</sup> We use an implementation of SVM that estimation of posterior class probabilities.

<sup>4</sup> <https://www.kaggle.com/lubaroli/notmnist/>



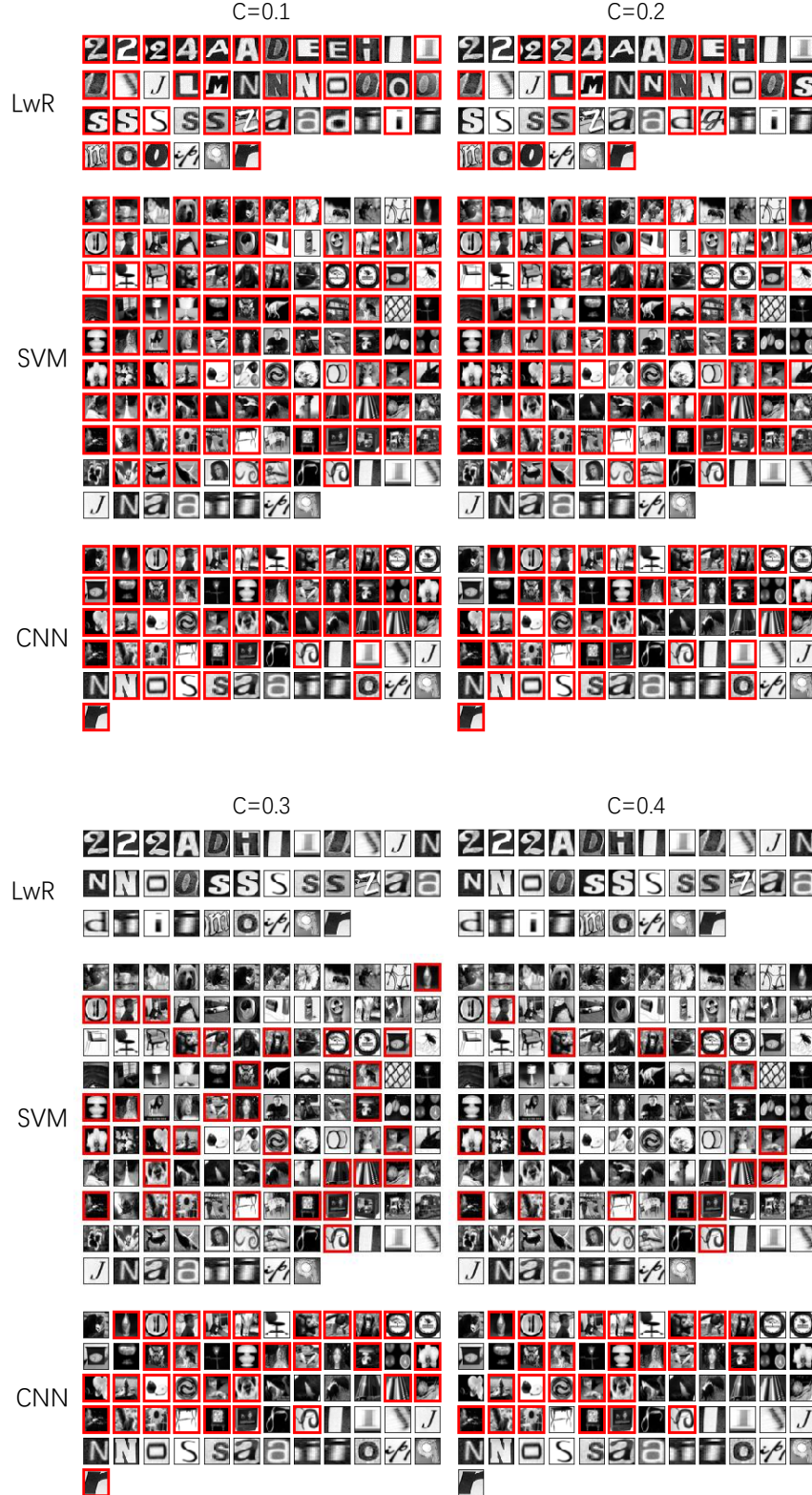
**Fig. 6.** Examples from Chars74k (upper row) and CIFAR100 (lower row) datasets.



**Fig. 7.** The relationship between accuracy and reject rate for character and non-character classification using Chars74k and CIFAR100.

notMNIST dataset. We also used 2,000 samples for CNN training, 2,000 samples for SVM and LwR training, and 2,000 samples for validations. We then obtain CNN classifier and SVM with reject option, and LwR. The CNN architecture in this experiment is as follows: two convolutional layers (3 kernels, kernel size  $3 \times 3$  with stride 1), two pooling layers (kernel size  $2 \times 2$  with stride 2) with Rectified Linear Units (ReLU) as the activation function, three fully-connected layers (3,136, 1,568, and 784 hidden units respectively), with minimization of cross-entropy loss.

Figure 5 shows the relationship between accuracy and reject rate, evaluated by using the test dataset. LwR clearly shows significantly better performance than the standard SVM with rejection option. Surprisingly, CNN could show slightly better performance than LwR at the stage when the rejection rate is low (although LwR showed the highest accuracy in this experimental range). This might be because the classifier of CNN, i.e., fully-connected layers of CNN is trained (or tuned) together with the feature extraction part of CNN and thus CNN has an advantage by itself. We will see that LwR outperforms in a more realistic experiment shown in the next section.



**Fig. 8.** Test samples misclassified by the LwR, the standard SVM and the CNN at different  $c \in \{0.1, 0.2, 0.3, 0.4\}$ . Rejected samples are highlighted in red.

### 4.3 Classification of Character and Non-character with Reject Option

As a more realistic application, we consider the classification of character and non-character. It is known that this classification task is a part of scene text detection task and still a tough recognition task (e.g., [38]) because of the ambiguity between characters and non-characters. In this experiment, as a scene character image dataset, we use Chars74k dataset<sup>5</sup>, which contains large number of character images in the natural scene. As a non-character image dataset, we use CIFAR100 dataset<sup>6</sup> which contains various kinds of objects in the natural scene. The image samples from those datasets were converted into gray scale and resized into  $32 \times 32$  pixels. Figure 6 shows several examples.

We randomly sample 5,000 images respectively from both CIFAR100 and Chars74k datasets. Among them, 6,000 are used for CNN training, 2,000 for SVM and LwR training, and 2,000 samples for validations. After training, we obtain CNN classifier and SVM with rejection thresholds, and LwR. CNN with the following architecture was used for this task: convolutional layers (32, 64, 128 kernels, kernel size  $3 \times 3$ ,  $5 \times 5$ ,  $5 \times 5$ , with stride 1, 2, 2 respectively, using ReLU as activation function), one max pooling layer ( $2 \times 2$  sized kernel with stride 2) and two fully-connected layers (512 and 2 units with ReLU and softmax), with minimization of cross entropy.

Figure 7 shows the test accuracy using 2,000 samples randomly chosen from Chars74k and CIFAR100<sup>7</sup>. We can see LwR could achieve the best compromise between accuracy and rejection rate among three methods. Surprisingly, when  $c = 0.3$  and  $c = 0.4$ , LwR predicts the labels of all test samples (i.e., no rejected test samples) with keeping a higher accuracy than CNN. This result confirm the strength of theoretically formulated LwR that obtains the classification and rejection function for minimizing the risk  $R$ , while the rejection option for CNN and SVM is rather heuristic.

Figure 8 shows the image samples are to be misrecognized without the rejection function. Namely, if we only rely on  $f(x)$ , the sample  $x$  is misrecognized. Among them, samples highlighted by red boxes are samples *successfully rejected* by the rejection function  $r(x)$ . Those highlighted samples represent that classification function  $f(x)$  and rejection function  $r(x)$  work in a complementary manner. In addition, we can find LwR has a lot of different rejected samples from SVM and CNN, whereas SVM and CNN share a lot of common samples. This result is induced by the effect that LwR employs different feature spaces for classification and rejection.

<sup>5</sup> <http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k/>

<sup>6</sup> <https://www.cs.toronto.edu/~kriz/cifar.html>

<sup>7</sup> CIFAR100 has much more samples for testing, but unfortunately, Chars74k remains highly limited number of samples for testing.

## 5 Conclusion

We propose an optimal rejection method for character recognition tasks. The rejection function is trained together with a classification function under the framework of Learning-with-Rejection (LwR). One technical highlight is that the rejection function can be trained on an arbitrary feature space which is different from the feature space for the classification. Another highlight is that LwR is not heuristic and its performance is guaranteed by a machine learning theory. From the application side, this is the first trial of using LwR for practical tasks. The experimental results show that the optimal rejection is useful for tough character recognition tasks than traditional strategies, such as SVM with a threshold. Since LwR is a general framework, we believe that it will be applicable to more tough recognition tasks and possible to achieve stable performance by its rejection function.

As future work, we plan to extend our framework to more realistic problem setting such as multi-class setting by using [39]. Furthermore, we will consider co-training framework of classification feature space and rejection feature space, which enables the truly optimal classification with reject option.

## Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP17H06100 and JP18K18001.

## Appendix

The optimal solution of (2) has a theoretical guarantee of the generalization performance as follows:

**Theorem 1 (proposed by [9]).** *Suppose that we choose classification function  $f$  from a function set  $F$  and choose rejection function  $r$  from a function set  $G$ . We denote  $R$  for test samples by  $R_{test}$ , and  $R$  for training samples by  $R_{train}$ . Then, the following holds:*

$$R_{test}(f, r) \leq R_{train}(f, r) + \mathfrak{R}(F) + (1 + c)\mathfrak{R}(G). \quad (3)$$

In above,  $\mathfrak{R}$  is the Rademacher complexity [40], which is a measure evaluating the theoretical overfitting risk of a function set. Roughly speaking, this theorem says that, if we prepare a proper feature space (i.e., not-so-complex feature space) mapped by  $\Phi$  and  $\Phi'$ , we will achieve a performance on test samples as high as on training samples. Thus, by solving the problem (2) of minimizing  $R_{train}$ , we can obtain theoretically optimal classification and rejection function which achieves the best balance between the classification error for not-rejected samples and the rejection cost [34].

## References

1. F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *CVPR*, 2015.
2. Y. Sun, D. Liang, X. Wang, and X. Tang, “Deepid3: Face recognition with very deep neural networks,” *arXiv preprint arXiv:1502.00873*, 2015.
3. S. S. Rautaray and A. Agrawal, “Vision based hand gesture recognition for human computer interaction: a survey,” *Artificial intelligence review*, 2015.
4. R. Rouhi, M. Jafari, S. Kasaei, and P. Keshavarzian, “Benign and malignant breast tumors classification based on region growing and cnn segmentation,” *Expert Systems with Applications*, 2015.
5. Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical attention networks for document classification,” in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016.
6. F. Condessa, J. Bioucas-Dias, and J. Kovačević, “Supervised hyperspectral image classification with rejection,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2016.
7. L. B. Marinho, J. S. Almeida, J. W. M. Souza, V. H. C. Albuquerque, and P. P. Rebouças Filho, “A novel mobile robot localization approach based on topological maps using classification with reject option in omnidirectional images,” *Expert Systems with Applications*, 2017.
8. P. L. Bartlett and M. H. Wegkamp, “Classification with a reject option using a hinge loss,” *Journal of Machine Learning Research*, 2008.
9. C. Cortes, G. DeSalvo, and M. Mohri, “Learning with Rejection,” in *ALT*, 2016.
10. M. A. Hernández and S. J. Stolfo, “Real-world Data is Dirty: Data Cleansing and the Merge/Purge Problem,” *DMKD*, 1998.
11. J. I. Maletic and A. Marcus, “Data Cleansing: Beyond Integrity Analysis.” in *IQ*, 2000.
12. Z. Khayyat, I. F. Ilyas, A. Jindal, S. Madden, M. Ouzzani, P. Papotti, J.-A. Quiané-Ruiz, N. Tang, and S. Yin, “Bigdancing: A System for Big Data Cleansing,” in *SIGMOD*, 2015.
13. T. M. Lehmann, C. Gonner, and K. Spitzer, “Survey: Interpolation Methods in Medical Image Processing,” *TMI*, 1999.
14. M. J. McAuliffe, F. M. Lalonde, D. McGarry, W. Gandler, K. Csaky, and B. L. Trus, “Medical Image Processing, Analysis and Visualization in Clinical Research,” in *CBMS*, 2001.
15. L. Neumann and J. Matas, “Efficient Scene Text Localization and Recognition with Local Character Refinement,” in *ICDAR*, 2015.
16. R. Messina and J. Louradour, “Segmentation-free Handwritten Chinese Text Recognition with LSTM-RNN,” in *ICDAR*, 2015.
17. K. Chen, M. Seuret, M. Liwicki, J. Hennebert, and R. Ingold, “Page Segmentation of Historical Document Images with Convolutional Autoencoders,” in *ICDAR*, 2015.
18. E. Rahm and H. H. Do, “Data Cleaning: Problems and Current Approaches,” *Data Engineering Bulletin*, 2000.
19. M. Mezzanzanica, R. Boselli, M. Cesarini, and F. Mercorio, “A Model-based Approach for Developing Data Cleansing Solutions,” *JDIQ*, 2015.

20. M. L. Lee, H. Lu, T. W. Ling, and Y. T. Ko, "Cleansing Data for Mining and Warehousing," in *DEXA*, 1999.
21. H. Chen, W.-S. Ku, H. Wang, and M.-T. Sun, "Leveraging Spatio-temporal Redundancy for RFID Data Cleansing," in *SIGMOD*, 2010.
22. D. Fecker, A. Asi, W. Pantke, V. Märgner, J. El-Sana, and T. Fingscheidt, "Document Writer Analysis with Rejection For Historical Arabic Manuscripts," in *ICFHR*, 2014.
23. U. Pal, N. Sharma, T. Wakabayashi, and F. Kimura, "Off-line Handwritten Character Recognition of Devnagari Script," in *ICDAR*, 2007.
24. Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, "Handwritten Digit Recognition with a Back-propagation Network," in *NIPS*, 1990.
25. S. D. Olabarriaga and A. W. Smeulders, "Interaction in the segmentation of medical images: A survey," *Medical image analysis*, 2001.
26. A. Mekhmoukh and K. Mokrani, "Improved fuzzy c-means based particle swarm optimization (ps) initialization and outlier rejection with level set methods for mr brain image segmentation," *Computer methods and programs in biomedicine*, 2015.
27. T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "PCANet: A Simple Deep Learning Baseline for Image Classification?" *TIP*, 2015.
28. J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained Linear Coding for Image Classification," in *CVPR*, 2010.
29. D. Cireşan, U. Meier, and J. Schmidhuber, "Multi-column Deep Neural Networks for Image Classification," *arXiv preprint arXiv:1202.2745*, 2012.
30. Y. Kessentini, T. Burger, and T. Paquet, "A Dempster-Shafer Theory Based Combination of Handwriting Recognition Systems with Multiple Rejection Strategies," *Pattern Recognition*, 2015.
31. D. P. Mesquita, L. S. Rocha, J. P. P. Gomes, and A. R. R. Neto, "Classification with Reject Option for Software Defect Prediction," *Applied Soft Computing*, 2016.
32. C. L. He, L. Lam, and C. Y. Suen, "A Novel Rejection Measurement in Handwritten Numeral Recognition Based on Linear Discriminant Analysis," in *ICDAR*, 2009.
33. X.-X. Niu and C. Y. Suen, "A Novel Hybrid CNN-SVM Classifier for Recognizing Handwritten Digits," *Pattern Recognition*, 2012.
34. D. S. Maitra, U. Bhattacharya, and S. K. Parui, "CNN Based Common Approach to Handwritten Character Recognition of Multiple Scripts," in *ICDAR*, 2015.
35. Y. Serdouk, H. Nemmour, and Y. Chibani, "New Off-line Handwritten Signature Verification Method Based on Artificial Immune Recognition System," *Expert Systems with Applications*, 2016.
36. R. Bertolami, M. Zimmermann, and H. Bunke, "Rejection Strategies for Offline Handwritten Text Line Recognition," *Pattern Recognition Letters*, 2006.
37. P. Li, L. Peng, and J. Wen, "Rejecting Character Recognition Errors Using CNN Based Confidence Estimation," *Chinese Journal of Electronics*, 2016.
38. X. Bai, B. Shi, C. Zhang, X. Cai, and L. Qi, "Text/Non-text Image Classification in the Wild with Convolutional Neural Networks," *Pattern Recognition*, 2017.
39. J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large Margin DAGs for Multi-class Classification," in *NIPS*, 2000.
40. P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian Complexities: Risk Bounds and Structural Results," *JMLR*, 2003.