# A Classification Performance Evaluation Measure Considering Data Separability

Lingyan Xue, Xinyu Zhang(✉), Weidong Jiang, Kai Huo, and Qinmu Shen

National University of Defense Technology, Changsha, China
zhangxinyu90111@163.com

**Abstract.** Machine learning and deep learning classification models are data-driven, and the model and the data jointly determine their classification performance. It is biased to evaluate the model's performance only based on the classifier accuracy while ignoring the data separability. Sometimes, the model exhibits excellent accuracy, which might be attributed to its testing on highly separable data. Most of the current studies on data separability measures are defined based on the distance between sample points, but this has been demonstrated to fail in several circumstances. In this paper, we propose a new separability measure– the rate of separability (RS), which is based on the data coding rate. We validate its effectiveness as a supplement to the separability measure by comparing it to four other distance-based measures on synthetic dataset. Then, we discover the positive correlation between the proposed measure and recognition accuracy in a multi-task scenario constructed from a real dataset. Finally, we discuss the methods for evaluating the classification performance of machine learning and deep learning models considering data separability.

**Keywords:** Machine learning · Classification accuracy · Data separability · Classification difficulty · Performance evaluation

## 1 Introduction

As an important branch in data mining, classification aims to construct a classification model to learn a mapping regularity from existing data to class labels. The research of model is essential, yet data also determines the performance [2]. A specific example is the impact of spectral separability on classification accuracy [17]. Numerous classification models have been proposed, including KNN, SVM, logistic regression, neural networks, etc., but studies on data separability are substantially fewer. A recent study in hyperspectral image classification has argued that insufficient data may limit the assessment capability of existing accuracy indexes [9]. That leads to the problem of whether a model performs best on a classification case is unclear or inconclusive [15]. It is acknowledged that a good classification model provides greater generalization potential, which means finding rules consistent with available data that apply widely to predict


L. Iliadis et al. (Eds.): ICANN 2023, LNCS 14254, pp. 1–13, 2023.
https://doi.org/10.1007/978-3-031-44207-0_1

the class of unknown data [20]. Yet the criteria for assessing the model's generalization ability remain debated. To simplify the performance evaluation process, researchers generally tend to adopt measures based on the confusion matrix [7], like accuracy, precision, kappa statistic, and F-score. Each measure is represented with a single score number, making it straightforward to compare and analyze classification models quantitatively. Although the result is intuitive, its comparability is invalid when confronted with a multi-task classification situation more representative of the real-world environment.

A contradictory example is that a classifier reaches the highest accuracy in one task but the lowest in another. What causes the problem is that such classifier-oriented measures treat the different instances of a dataset as statistical objects and ignore the classification difficulty of each instance. For the above issue, Yu et al. [18] proposed an instance-oriented measure but only apply to data with few samples due to the computational complexity of classification difficulty for each instance. Therefore, we require a measure to statistically characterize the classification difficulty of datasets. Fortunately, previous research has established that separability is an intrinsic characteristic of a dataset [4] to describe how instances belonging to different classes mix. Measuring the data quality is critical for estimating the problem's difficulty in advance since a classification model's accuracy strongly depends on the data quality [1]. Obviously, the more separable the dataset, the simpler the classification. Eventually, we consider data separability as a metric of classification difficulty.

There are several measures of data separability that can quantify classification difficulty. The Fisher discriminant ratio [8] has been used in many studies, which measures the data separability using the mean and standard deviation of each class, but it fails in some cases like a two-class circle data. A more effective issue is data complexity which measures the distance of intra classes as well as the inter class. Ho and Basu [6] conducted a groundbreaking review of data complexity measures. Recently, Lorena et al. [11] summarized existing methods for the measurement of classification complexity, showing that some of those may have large time cost.

As an alternative to the distance-based criterion, we consider explaining data separability from the perspective of probability theory. Inspired by Cover and Thomas [3], the process of minimizing the data rate distortion is equivalent to the process of solving the optimal solution of the likelihood function, i.e., the data coding rate has strong consistency with the parameter estimation performance [13]. That means if the data can be fitted with better distribution model after segmentation, then the data should be effectively encoded in relation to such model. Ma et al. [12] argued that the coding rate (subject to a distortion) provides a natural measure of the goodness of segmentation for real-valued mixed data.

Since there is no research verifying the feasibility of using coding rate as a measure of data separability, this is the first study to construct a separability measure based on rate-distortion theory called the rate of separability (RS). The main contributions of this paper are summarized as follows.

1) We propose a data separability measure based on rate-distortion theory, and verify its effectiveness in theory and experiments.

2) We find a positive correlation between classification accuracy and data separability in a multi-task noisy environment.

3) In a multi-task noisy environment, we design a task-oriented classifier performance evaluation method considering data separability as the task difficulty. Unlike the classification accuracy changing with different tasks, this method obtains classifier ability as the classifier's inherent property under certain assumptions.

4) We build a modular classifier performance evaluation model to explain the function of deep learning convolutional blocks using data separability.

The rest of this paper is structured as follows. Section 2 introduces the method of constructing coding-rate-based measure. Section 3 provides experimental methods for validating measure validity and evaluates the classification model performance; results and analysis are also given in this section. Finally, we conclude in Sect. 4.

## 2 Data Separability Measure

In this section, we apply rate-distortion theory in constructing a new data separability measure.

### 2.1 Coding-Rate Based Data Separability Measure

Given a data $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_m] \in \mathbb{R}^{d\times m}$ with $m$ samples of $d$ dimension and a encoding precision $\varepsilon > 0$, let $\mathbf{\Pi} = \{\mathbf{\Pi}^j \in \mathbb{R}^{m\times m}\}_{j=1}^{k}$ be the label matrix of the $\mathbf{X}$ in the $k$ classes, and $\mathbf{\Pi}^j(i,i)$ is the label of $\mathbf{x}_i$ belonging to class $j$, our proposed data separability measure based on rate-distortion is:

$$\mathrm{R_S}(\mathbf{X}) = \frac{\mathrm{R_C}(\mathbf{X}, \varepsilon|\mathbf{\Pi})}{\mathrm{R}(\mathbf{X}, \varepsilon)}. \tag{1}$$

In Eq. (1), the $\mathrm{R_C}(\mathbf{X}, \varepsilon|\mathbf{\Pi})$ and $\mathrm{R}(\mathbf{X}, \varepsilon)$ denote the local and global coding rate of the data, respectively. Unlike Yu et al. [19], who utilized $\Delta\mathrm{R}(\mathbf{X}) = \mathrm{R}(\mathbf{X}, \varepsilon) - \mathrm{R_C}(\mathbf{X}, \varepsilon|\mathbf{\Pi})$ as the optimization problem's objective function subjecting to $\left\|\mathbf{X}^j\right\|_F^2 = \mathrm{tr}(\mathbf{\Pi}^j)$, here we discard the constraint and adopt a ratio form between $\mathrm{R}(\mathbf{X}, \varepsilon|\mathbf{\Pi})$ and $\mathrm{R}(\mathbf{X}, \varepsilon)$, resulting in a data separability measure $\mathrm{R_S}(\mathbf{X})$ in the range of [0,1] with low values indicating high separability. It means that the smaller the $\mathrm{R_C}(\mathbf{X}, \varepsilon|\mathbf{\Pi})$, the more clustered the samples within the class, and the larger the $\mathrm{R}(\mathbf{X}, \varepsilon)$, the more dispersed the samples between classes. Next we introduce the definition of the coding rate and explain how the measure we proposed reflects the data intrinsic separability.

### 2.2 Definition and Computation of the Coding Rate

According to Cover and Thomas' [3] definition of rate-distortion: the rate-distortion $\mathrm{R}(\mathbf{X}, \varepsilon)$ is the minimal number of binary bits needed to encode $\mathbf{X}$ and the expected decoding error is less than $\varepsilon$. The actual estimation coding rate of $\mathbf{X}$ with zero mean is as follows:

$$\mathrm{R}(\mathbf{X}, \varepsilon) = \frac{m}{2} \log\det(\mathbf{I} + \frac{d}{m\varepsilon^2}\mathbf{X}\mathbf{X}^T). \tag{2}$$

Furthermore, suppose $\mathbf{X}$ has $k$-class samples, then $\mathbf{X} = \mathbf{X}^1 \cup \mathbf{X}^2 \cup ... \cup \mathbf{X}^k$ . the data $\mathbf{X}^j$ in each class $j$ also occupy a certain volume in its low dimensional subspace. For each subset, the above coding rate (2) is applied, then $\mathrm{R_C}(\mathbf{X}, \varepsilon|\mathbf{\Pi})$ is given by

$$\mathrm{R}_C(\mathbf{X}, \varepsilon|\mathbf{\Pi}) = \sum_{j=1}^{k} \frac{tr(\mathbf{\Pi}^j)}{2} \log\det\left(\mathbf{I} + \frac{d}{tr(\mathbf{\Pi}^j)\varepsilon^2}\mathbf{X}\mathbf{\Pi}^j\mathbf{X}^T\right) \tag{3}$$

The equation for the coding rate in Eq. (2) is for the scenario where the mean value of the given data is zero mean. More generally, when $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_m] \in \mathbb{R}^{d\times m}$ is not zero mean, we have the mean $\boldsymbol{\mu} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i \in \mathbb{R}^d$ and the zero mean part of the data $\bar{\mathbf{X}}$, thus the total coding rate of $\mathbf{X}$ with non-zero mean is:

$$\mathrm{R}(\mathbf{X}) = \frac{m}{2} \log\det(\mathbf{I} + \frac{d}{m\varepsilon^2}\bar{\mathbf{X}}\bar{\mathbf{X}}^T) + \frac{d}{2}\log_2\left(1 + \frac{\boldsymbol{\mu}^T\boldsymbol{\mu}}{\varepsilon^2}\right). \tag{4}$$

### 2.3 Correlation Between RS and Data Separability

This section discusses the connection between RS and data separability. Under the condition that the data follows a Gaussian distribution, we prove Theorem 1. Theorem 1 gives the lower bound of the data coding rate and the necessary and sufficient conditions for it to reach the lower bound. This condition illustrates that if and only if every class $\mathbf{X}$ has the same distribution, the total coding rate of $\mathbf{X}$ is identical to the sum of $\mathbf{X^j}$ 's coding rate.

**Theorem 1.** *For any* $\left\{\mathbf{X}^j \in \mathbb{R}^{d\times m_j}\right\}_{j=1}^{k}$ *and any* $\varepsilon > 0$, *let* $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_m] = \left[\mathbf{X}^1, \cdots \mathbf{X}^k\right] \in \mathbb{R}^{d\times m}$ *with* $m = \sum_{j=1}^{k} m_j$ *and* $\boldsymbol{\mu} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i \in \mathbb{R}^d$, *then we define the zero mean part* $\bar{\mathbf{X}} = \mathbf{X} - \boldsymbol{\mu}\cdot\mathbf{1}_{1\times m}$. *Let* $\mathbf{X}^j = [\mathbf{x}_1^j, ..., \mathbf{x}_{m_j}^j] \in \mathbb{R}^{d\times m_j}$ *with* $\boldsymbol{\mu}^j = \frac{1}{m_j}\sum_{i=1}^{m_j}\mathbf{x}_i^j \in \mathbb{R}^d$ *and* $\bar{\mathbf{X}}^j = \mathbf{X}^j - \boldsymbol{\mu}^j\cdot\mathbf{1}_{1\times m_j}$. *We have* $\mathrm{R}(\mathbf{X}, \varepsilon) \geq \mathrm{R_C}(\mathbf{X}, \varepsilon|\mathbf{\Pi})$,

$$\begin{aligned} &\frac{m}{2} \log\det(\mathbf{I} + \frac{d}{m\varepsilon^2}\bar{\mathbf{X}}\bar{\mathbf{X}}^T) + \frac{d}{2}\log_2\left(1 + \frac{\boldsymbol{\mu}^T\boldsymbol{\mu}}{\varepsilon^2}\right) \geq \\ &\sum_{j=1}^{k} \frac{m_j}{2} \log\det(\mathbf{I} + \frac{d}{m_j\varepsilon^2}\bar{\mathbf{X}}^j(\bar{\mathbf{X}}^j)^T) + \frac{d}{2k}\log_2\left(1 + \frac{(\boldsymbol{\mu}^j)^T\boldsymbol{\mu}^j}{\varepsilon^2}\right). \end{aligned} \tag{5}$$

*where the equality holds if and only if*

$$\frac{\bar{\mathbf{X}}^1(\bar{\mathbf{X}}^1)^T}{m_1} = \frac{\bar{\mathbf{X}}^2(\bar{\mathbf{X}}^2)^T}{m_2} = \cdots = \frac{\bar{\mathbf{X}}^k(\bar{\mathbf{X}}^k)^T}{m_k} = \frac{\bar{\mathbf{X}}(\bar{\mathbf{X}})^T}{m}$$
$$\boldsymbol{\mu}^1 = \boldsymbol{\mu}^2 = \cdots = \boldsymbol{\mu}^k = \boldsymbol{\mu}. \tag{6}$$

The Proof of Theorem 1 is based on the concave property of the $\log\det(\cdot)$ and $\log(\cdot)$ functions, and they satisfy Jensen's inequality.

*Proof.* Since $\log\det(\cdot)$ and $\log(\cdot)$ is strictly concave, The Jensen's inequality is satisfied. We have

$$f(\sum_{j=1}^{k}\beta_j\mathbf{S}^j) \geq \sum_{j=1}^{k}\beta_j f(\mathbf{S}^j). \tag{7}$$

for all $\{\beta_j > 0\}_{j=1}^{k}$ , $\sum_{j=1}^{k}\beta_j = 1$ and $\left\{\mathbf{S}^j \in \mathbb{S}_{++}^n\right\}_{j=1}^{k}$, where equality holds if and only if $\mathbf{S}^1 = \mathbf{S}^2 = \cdots = \mathbf{S}^k$ .

For function $\log\det(\cdot)$, take $\beta^j = \frac{m_j}{m}$ and $\mathbf{S}^j = \mathbf{I} + \frac{d}{m_j\varepsilon^2}\bar{\mathbf{X}}^j(\bar{\mathbf{X}}^j)^T$, we get

$$\log\det(\mathbf{I} + \frac{d}{m\varepsilon^2}\bar{\mathbf{X}}\bar{\mathbf{X}}^T) \geq \sum_{j=1}^{k}\frac{m_j}{m}\log\det(\mathbf{I} + \frac{d}{m_j\varepsilon^2}\bar{\mathbf{X}}^j(\bar{\mathbf{X}}^j)^T). \tag{8}$$

with equality holds if and only if $\frac{\bar{\mathbf{X}}^1(\bar{\mathbf{X}}^1)^T}{m_1} = \frac{\bar{\mathbf{X}}^2(\bar{\mathbf{X}}^2)^T}{m_2} = \cdots = \frac{\bar{\mathbf{X}}^k(\bar{\mathbf{X}}^k)^T}{m_k} = \frac{\bar{\mathbf{X}}(\bar{\mathbf{X}})^T}{m}$.

For function $\log(\cdot)$, take $\beta^j = \frac{1}{k}$ and $\mathbf{S}^j = 1 + \frac{d}{\varepsilon^2}(\boldsymbol{\mu}^j)^T\boldsymbol{\mu}^j$, we get

$$\log(1 + \frac{\boldsymbol{\mu}^T\boldsymbol{\mu}}{\varepsilon^2}) \geq \sum_{j=1}^{k}\frac{1}{k}\log(1 + \frac{(\boldsymbol{\mu}^j)^T\boldsymbol{\mu}^j}{\varepsilon^2}). \tag{9}$$

with equality holds if and only if $\boldsymbol{\mu}^1 = \boldsymbol{\mu}^2 = \cdots = \boldsymbol{\mu}^k = \boldsymbol{\mu}$, and the last equality is from $\sum_{j=1}^{k} m_j\boldsymbol{\mu}^j = m\boldsymbol{\mu}$. From formula (8) and (9), Theorem 1 can be proved.

From Theorem 1, we can conclude that the sum of the various classes of data coding rate is a lower bound on the overall data coding rate. When the overall data coding rate reaches the lower bound, its necessary and sufficient condition indicate: for the Gaussian distributed data, each category of data has the same distribution, which also means that the feature vectors of each class have a high degree of coincidence, corresponding to the most inseparable situation.

## 3 Experiments

### 3.1 Validation on Two-Class Synthetic Datasets

We first verify the proposed measure RS's effectiveness using a two-class synthetic dataset[1] with adjustable separability, and contrast its separability evalua-

[1] The datasets are created by the Samples Generator in sklearn.datasets https://scikit-learn.org/stable/modules/classes.html#samples-generator.

tion results with distance-based measures [5] (e.g., DSI, N2, LSC, Density). We experiment on the data following a Gaussian distribution, the region of feature overlap can be adjusted by changing the feature standard deviation (SD). We set the SD parameter from 1 to 9. Four instances are depicted in Fig. 1, the results are presented in Fig. 2.

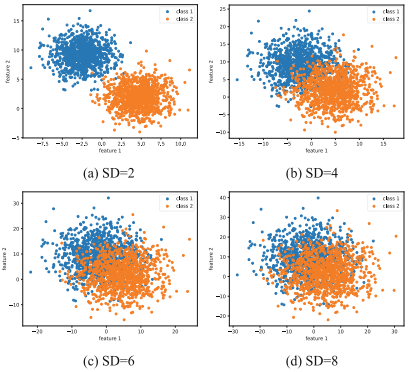


**Fig. 1.** The data with different cluster standard deviations (SD). A high SD value denotes a significant overlap area

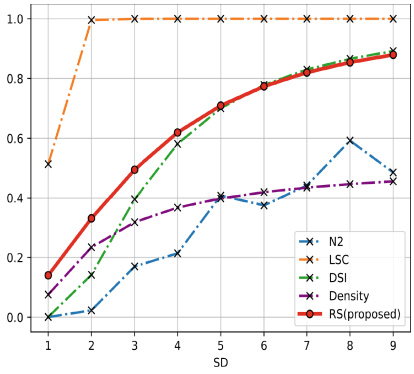


**Fig. 2.** Comparison of data separability evaluation results with varying degrees of feature overlap. For measures, a high value on the y-axis indicates low separability

In this condition, both N2 and LSC fail to assess the data separability. Among them, LSC can only distinguish the case of features with or without overlap and is not sensitive to the change of feature overlap area. At the same time, N2 fluctuates with the deterioration of data separability, suggesting a lower evaluation precision. Besides, RS, DSI, and Density can correctly reflect the trend of data separability, i.e., a high SD value corresponds to a high measure value. Furthermore, both DSI and RS have an extensive dynamic change range.

### 3.2 Correlation Between Classification Accuracy and Data Separability

After verifying the validity of RS as a measure of data separability, we characterize the data separability using RS values. This section discusses the experimental procedures used to investigate the correlation between classification accuracy and data separability. The experiment framework is shown in Fig. 3.

Step I is to add Gaussian white noise with a specific variance to the original data to create a test set with a signal-to-noise ratio (SNR) of 5–20 dB.

In Step II, we deploy four standard machine learning classifiers. Nonlinear classifiers such as K-Nearest Neighbor (KNN) and Support Vector Machine with

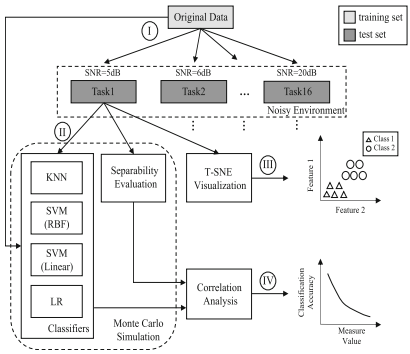


**Fig. 3.** Experiment procedure to verify the correlation between classification accuracy and data separability

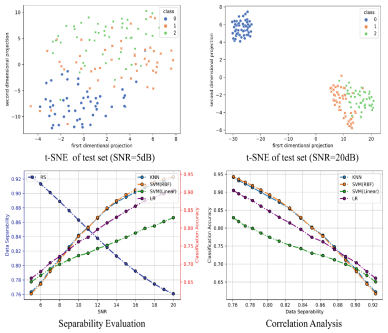


**Fig. 4.** Analysis results to verify the correlation between classification accuracy and data separability

Radial Basis Function (SVM with RBF) can generate nonlinear decision boundaries. Linear classifiers include linear SVM and logistic regression (LR). The Iris data from UCI repository [10] are utilized in the experiment.

In step III, the T-SNE tool is used to visualize the influence of noise on data separability. And in step IV, we compute noisy data's RS value and analyze the classification accuracy correlation. Here we apply the Monte Carlo simulation to average the randomness of the results due to noise.

The analysis results on the Iris data are shown in Fig. 4. When the SNR is lower than 5 dB, the test data separability becomes extremely poor relative to the training data. And when the SNR is high as 20dB, its separability is equivalent to the training data. Referring to the separability evaluation and correlation analysis, as SNR grows, the separability of test data gradually improves, and the classification accuracy increases along with it. Thus we can conclude a positive correlation between data separability and recognition accuracy.

### 3.3 Classifier's Ability Evaluated by Classification Accuracy Under Data Separability

In this section, we evaluate the classifier's generalization ability in the group of tasks constricted in Sect. 3.2. Specifically, as shown in Fig. 4, for the dataset Iris, the classification accuracy of SVM (RBF) is consistently higher than other classifiers at SNR = 20 dB, but the lowest at SNR = 5 dB. Since the 5 dB task is more difficult than the 20 dB one, we can't conclude whether the classifier performance is good or not. At this point, how could the classifier's performance be measured?

The simple idea is to assign a certain weight $\mathbf{W} \in \mathbb{R}^n$ to the recognition accuracy $\mathbf{P}_{acc} \in \mathbb{R}^n$ of the classifier on that group of tasks according to the

difficulty of the recognition task, and $n$ is the number of tasks. The classification ability $\theta$ on these tasks is defined as

$$\theta = \mathbf{W}^T \mathbf{P}_{acc}. \tag{10}$$

$\mathbf{W}$ is determined by the difficulty of the recognition task. The more difficult the task, the higher the weight value. According to the prior experiment, the task difficulty depends to some extent on the data separability. Thus, $\mathbf{W}$ as a mapping matrix is parameterized by the separability $\mathbf{R}_{\mathrm{S}}$. To quantify this mapping relationship, we seek a functional form $f(\cdot)$ of the mapping matrix $\mathbf{W}$.

$$\theta = f(P_{acc}; R_{\mathrm{S}}) \tag{11}$$

$f(\cdot)$ needs to be obtained by fitting a given $P_{acc}$ and $\theta$. $P_{acc}$ can be derived directly from the classification results, whereas $\theta$ is uncertain. Therefore, it is first necessary to construct the known $\theta$ based on the following assumptions.

1) For different difficulty tasks, homogeneous classifiers with fixed parameters exhibit different recognition accuracies.

2) For the same dataset, homogeneous classifiers with fixed parameters exhibit consistent recognition ability values.

3) For homogeneous classifiers with different parameter settings, their relative ability value can be inferred from the recognition accuracy.

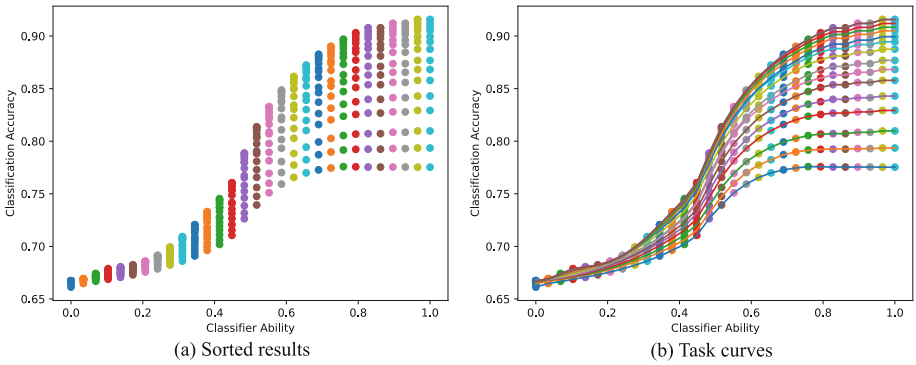


**Fig. 5.** The map of $\mathbf{P}_{\mathrm{acc}}$ and $\boldsymbol{\theta}$. Here, $k = 30$, $n = 15$. Each column of $\mathbf{P}_{\mathrm{acc}}$ records the classification accuracy of an SVM model on 16 tasks associated with the same color point column in Figure (a). Each row of $\mathbf{P}_{\mathrm{acc}}$ records the classification accuracy of 30 SVM models on a single task. The row values are sequentially concatenated to obtain the task curve shown in Figure (b)

Based on the assumptions stated above, we choose the SVM (Linear) model on the Iris dataset to perform the anti-noise experiment depicted in Fig. 3. $k$ SVMs with relative ability values $\boldsymbol{\theta}$ were obtained by adjusting the regular parameter $C$. $\boldsymbol{\theta} \in \mathbb{R}^k$ take $k$ values evenly from 0 to 1. Each SVM tests on $n$ noisy

tasks, and get $\mathbf{P}_{\text{acc}}^{j} \in \mathbb{R}^{n}(j = 1, 2, ..., k)$. $k$-group $\mathbf{P}_{\text{acc}}^{j}$ is sorted from small to large according to its largest element, and we have $\mathbf{P}_{\text{acc}} = [\mathbf{P}_{\text{acc}}^{1}, \mathbf{P}_{\text{acc}}^{2}, ..., \mathbf{P}_{\text{acc}}^{k}] \in \mathbb{R}^{n \times k}$. Figure 5 shows the mapping of $\mathbf{P}_{\text{acc}}$ and $\boldsymbol{\theta}$.

Observe that the shape of the curve in Fig. 5(b) is more consistent with that of the Sigmoid function, but the upper and lower bounds of the task curve are variable; thus, Eq. (12) is adopted as the fitting function.

$$P_{acc} = \frac{u - l}{1 + \exp(-a * (\theta - b))} + u \tag{12}$$

Plotting the curve of Eq. (12) in Fig. 6, we explore the properties of this function.

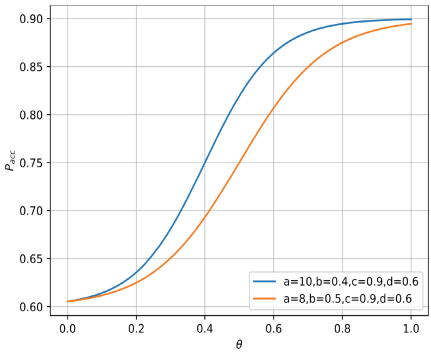


**Fig. 6.** Fitting function plot

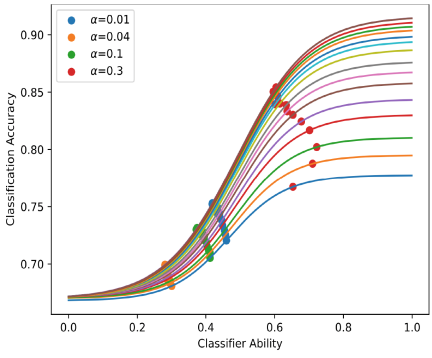


**Fig. 7.** The fitting task curves and the mapping points

The parameters $u$, $l$, $a$, and $b$ reflect the function properties as follows.

1) $u$ and $l$ can represent the classifier's upper and lower bounds of recognition accuracy on a set of recognition tasks, respectively.

2) $(u - l) * a$ reflects the slope of the function. The flatter the function, the harder the task and the lower the recognition accuracy.

3) $b$ affects the right shift rate of the function. The larger the right shift magnitude, the more difficult the task is, and the less accurate the classifier is.

Then we employ the polynomial fitting approach, with $a$ and $b$ represented by $R_{\text{S}}$

$$\begin{aligned} f_a(R_{\text{S}}) &= h_0 + h_1 R_{\text{S}} + h_2 R_{\text{S}}^2 \\ f_b(R_{\text{S}}) &= p_0 + p_1 R_{\text{S}} + p_2 R_{\text{S}}^2. \end{aligned} \tag{13}$$

The mapping function $f^{-1}(\cdot)$ from classification accuracy to classifier ability with separability $R_{\text{S}}$ as a parameter is now obtained.

$$P_{acc} = f^{-1}(\theta; R_S) = \frac{u - l}{1 + \exp(-f_a(R_S) * (\theta - f_b(R_S)))} + u \tag{14}$$

To examine the validity of this mapping function, we need to substitute the recognition accuracy of another classifier into the Eq. (14) to ensure the uniqueness of its recognition ability value, demonstrating that the recognition ability value exists as an inherent property of the classifier.

Figure 7 shows the results of fitting the task curve with the SVM model as a reference and an evaluation of the LR (with adjusted parameter $\alpha$) models' recognition ability on this curve.

This evaluation method has a high assessment accuracy in the middle of the task curve. The evaluation of the LR model ability values for $\alpha = 0.01$ and $\alpha = 0.1$ are distributed over a small interval, and a set of recognition accuracies essentially map to a unique recognition ability value. Whereas at the two ends of the curve, a slight change in recognition accuracy may bring about a significant deviation in recognition ability due to the presence of the saturation zone.

### 3.4 CNN Layers' Performance Evaluated by Data Separability

As an extension of machine learning classifiers, deep learning classifiers have greatly improved recognition performance but are not satisfactory in model interpretability. The convolution module, for example, is widely believed to play the role of feature extraction, enabling the final output data to be more separable, but how do we measure this function? In this section, we design a modular classifier recognition performance evaluation method to evaluate the performance of each convolutional component of the CNN.

Effective separability indices are invaluable for the performance evaluation of radar signal classification algorithms [14]. Since radar image is more difficult to identify the classes to which they belong after their semantic features are extracted by the convolutional layer, we use the typical radar image MSTAR[2] as the experimental data. The evaluation method proposed is to insert a feature separability analysis module after each convolutional block to monitor its performance, and the separability measure used is RS.

On the MSTAR dataset, we evaluate the performance of some convolutional blocks of CNN provided by Chen et al. [16]. And to reduce the computational effort of RS, we apply a $2 * 2$ average pooling to the feature map. The network structure and the feature separability analysis module are shown in Fig. 8.

After 100 epochs of training, the convergence of recognition accuracy on the test set and the variation of feature separability extracted by each convolutional block are set out in Fig. 9.

The most striking result from Fig. 9 is that the separability of features extracted by each convolutional block keeps step with the final classification accuracy. When classification accuracy improves dramatically, the RS value falls

[2] The URL for downloading the dataset: https://www.sdms.afrl.af.mil/datasets/mstar/.

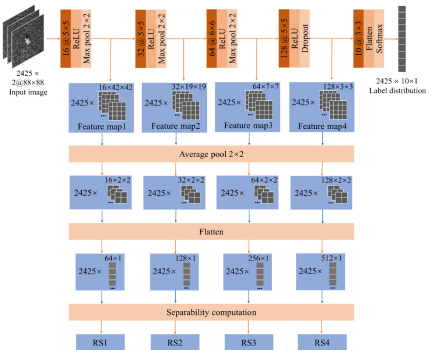


**Fig. 8.** Network architecture and feature separability analysis module

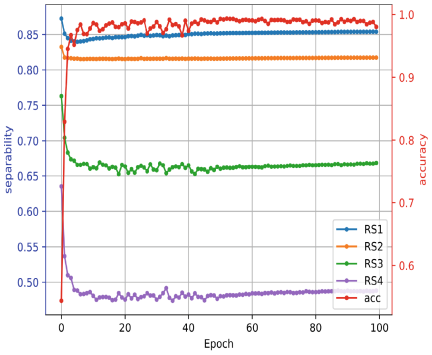


**Fig. 9.** Classification accuracy and feature separability analysis results

precipitously. And when the classification accuracy converges, the RS value becomes steady. Furthermore, we also find that the convolutional block in a deeper network has a more significant function in extracting a more separable feature with a lower RS value (RS1 > RS2 > RS3 > RS4). And the deeper feature map exhibits a wider dynamic range of RS value ($\Delta$RS1 < $\Delta$RS2 < $\Delta$RS3 < $\Delta$RS4).

## 4 Conclusion

Data separability quantification provides some basis for analyzing, understanding, and enhancing model performance. In this paper, we validate the effectiveness of the proposed measure on a typical synthetic two-class dataset and confirm its positive correlation with the classification accuracy in a series of noisy tasks constructed from real datasets. Then we designed machine learning and deep learning classifier model evaluation methods based on the above two basic argumentation experiments. We build a functional mapping model for machine learning classifiers from classification accuracy to classifier ability. In the model, the task difficulty is characterized by the measure, and the classification accuracy assesses the classifiers' capability value as its inherent properties with a separability measure as a parameter. For deep learning classifiers, we use a modular evaluation approach. Each convolutional block's ability to extract separable features is assessed using the proposed measure. Finally, we explain why neural networks work effectively from the perspective of feature separability.

In fact, the separability measure can also be applied to evaluate clustering results, understand the demerit of each feature, provide a theory for building multi-classifier decisions, or reduce data complexity as a loss function. In general, explaining and improving classification performance by exploiting data separability still deserves further study.

**Acknowledgments.** This work was supported in part by Hunan Provincial Natural Science Foundation of China under Grants 2021JJ20056 and National Natural Science Foundation of China under Grants 61921001.

## References

1. Bello, M., Nápoles, G., Vanhoof, K., Bello, R.: Data quality measures based on granular computing for multi-label classification. Inf. Sci. **560**, 51–67 (2021). https://doi.org/10.1016/j.ins.2021.01.027
2. Cano, J.R.: Analysis of data complexity measures for classification. Expert Syst. Appl. **40**(12), 4820–4831 (2013). https://doi.org/10.1016/j.eswa.2013.02.025
3. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley Interscience, New York (2006)
4. Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F.: Data intrinsic characteristics. In: Learning from Imbalanced Data Sets, pp. 253–277. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98074-4_10
5. Guan, S., Loew, M.: A novel intrinsic measure of data separability (2022). https://doi.org/10.1007/s10489-022-03395-6
6. Ho, T.K., Basu, M.: Complexity measures of supervised classification problems. IEEE Trans. Pattern Anal. Mach. Intell. **24**(3), 289–300 (2002). https://doi.org/10.1109/34.990132
7. Hossin, M., Sulaiman, M.N.: A review on evaluation metrics for data classification evaluations. Int. J. Data Min. Knowl. Manage. Process **5**(2), 1–11 (2015). https://doi.org/10.5121/ijdkp.2015.5201
8. Li, C., Wang, B.: Fisher linear discriminant analysis. CCIS Northeastern University (2014)
9. Li, S., Hao, Q., Gao, G., Kang, X.: The effect of ground truth on performance evaluation of hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. **56**(12), 7195–7206 (2018). https://doi.org/10.1109/TGRS.2018.2849225
10. Lichman, M.E.A.: UCI machine learning repository (2013). https://archive.ics.uci.edu/ml/datasets.php
11. Lorena, A.C., Garcia, L.P.F., Lehmann, J., Souto, M.C.P., Ho, T.K.: How complex is your classification problem?: A survey on measuring classification complexity. ACM Comput. Surv. **52**(5), 1–34 (2019). https://doi.org/10.1145/3347711
12. Ma, Y., Derksen, H., Hong, W.: Segmentation of multivariate mixed data via lossy data coding and compression. IEEE Trans. Pattern Anal. Mach. Intell. **29**(9), 1546–1562 (2007). https://doi.org/10.1109/TPAMI.2007.1085
13. Madiman, M., Harrison, M., Kontoyiannis, I.: Minimum description length versus maximum likelihood in lossy data compression. In: International Symposium on Information Theory. IEEE, Chicago (2004). https://doi.org/10.1109/ISIT.2004.1365499
14. Mishra, A.K.: Separability indices and their use in radar signal based target recognition. IEICE Electron. Express **6**(14), 1000–1005 (2009). https://doi.org/10.1587/elex.6.1000
15. Oprea, M.: A general framework and guidelines for benchmarking computational intelligence algorithms applied to forecasting problems derived from an application domain-oriented survey. Appl. Soft Comput. **89**, 106103 (2020). https://doi.org/10.1016/j.asoc.2020.106103

16. Sizhe, C., Haipeng, W., Feng, X., Yaqiu, J.: Target classification using the deep convolutional networks for SAR images. IEEE Trans. Geosci. Remote Sens. **54**(8), 4806–4817 (2016). https://doi.org/10.1109/TGRS.2016.2551720
17. Wicaksonoa, P., Aryagunab, P.A.: Analyses of inter-class spectral separability and classification accuracy of benthic habitat mapping using multispectral image. Remote Sens. Appl. Soc. Environ. **19**, 100335 (2020). https://doi.org/10.1016/j.rsase.2020.100335
18. Yu, S., Li, X., Feng, Y., Zhang, X., Chen, S.: An instance-oriented performance measure for classification. Inf. Sci. **580**, 598–619 (2021). https://doi.org/10.1016/j.ins.2021.08.094
19. Yu, Y., Chan, K.H.R., You, C., Song, C., Ma, Y.: Learning diverse and discriminative representations via the principle of maximal coding rate reduction (2020). https://doi.org/10.48550/arXiv.2006.08558
20. Zhang, C., Samy, B., Moritz, H., Benjamin, R., Oriol, V.: Understanding deep learning (still) requires rethinking generalization. Commun. ACM **64**(3), 107–115 (2021). https://doi.org/10.1145/3446776