Journal of Mathematical Modeling
Vol. 13, No. 4, 2025, pp. 875-885. Research Article

JMM

Is evaluation based on accuracy of classification algorithms misleading? An approach to model validation using Bayes error rate

Hossein Kazemzadeh Gharechopogh and Adel Mohammadpour*

Faculty of Mathematics and Computer Science, Amirkabir University of Technology
Email(s): kazemzadeh_hk@aut.ac.ir, adel@aut.ac.ir

Abstract. Researchers have long regarded model accuracy as the primary metric for evaluating the performance of classification algorithms. The current evaluation approach, which relies solely on model accuracy, often leads to inappropriate evaluation of classifiers, regardless of the dataset's separability and complexity. This limitation underscores the need for a new and more comprehensive method. We argue that accuracy-based evaluation can be misleading, even when considering measures of data separability and complexity. We compare the error rates of well-known classifiers on Gaussian-generated datasets and show that, paradoxically, many algorithms' observed errors are lower than that of the theoretical optimal classifier, leading to an overestimation of their performance. We consider a model invalid if its error rate is lower than the optimal classifier error, known as the Bayes error rate. To identify such invalid models, we introduce a procedure and propose an algorithm for model validation based on the Bayes error rate.

Keywords: Classification, evaluation, validation, Bayes error rate, discriminant analysis and complexity measure.
AMS Subject Classification 2020: 62R07.

1 Introduction

Model evaluation is the process of finding the optimal classifier for prediction. How do we evaluate and compare classification models? Researchers typically address this question by comparing the accuracy of different algorithms [12]. While the classifiers' predictive accuracy reflects the performance of the algorithms, a fair evaluation also necessitates considering the complexity of the classification task.

*Corresponding author
Received: 29 January 2025 / Revised: 24 February 2025 / Accepted: 4 March 2025
Special Issue: First International Conference on Machine Learning and Knowledge Discovery, Amirkabir University of Technology-Tehran Polytechnic, December 18-19, 2024
DOI: 10.22124/jmm.2025.28746.2649

As an operator, a classifier takes a vector of variables (features) and produces an output decision that determines its label. The efficiency of algorithms in solving classification problems is defined by their error rates. Moreover, the performance of a classifier is influenced by both the efficiency of the algorithms and the complexity (or separability) of the datasets [7].

The concept of complexity in classification problems was introduced in Ho and Basu's work while analyzing the difficulties of classification problems [8]. They proposed a measure to describe the complexity of binary classification problems based on the geometrical complexity of the class boundary. They categorized proposed measures into three main groups: measures of overlap of features, class separability, and density (topology and geometry) measure. After that, the complexity measure is redefined into six characterized measures: feature-based, linearity, neighborhood, network, dimensionality, and class imbalance [10].

Most researchers ignore the complexity measures to simplify the evaluation process and focus only on some items according to the confusion matrix [1]. The four main criteria often used in evaluation are accuracy, sensitivity, precision, and specificity; however, others may also be used [2]. Several libraries and packages have been developed in some programming languages to compute these measures. One of the developed packages for model evaluation is HungaBunga [15]. This package employs a brute-force approach to rank all scikit-learn models by accuracy, tuning all possible hyperparameters in the process.

Some researchers contend that this evaluation approach can lead to inappropriate decisions if the separation characteristics of the data are ignored [18]. They illustrate this concern with a paradoxical example where a classifier achieves the highest accuracy on one dataset but performs poorly on another. They argue that measures of complexity, such as separability, are intrinsic characteristics of a dataset [5]. Consequently, they categorized classification tasks based on the complexity of each dataset and proposed instance-oriented measures for evaluation [19]. However, the computational complexity involved in assessing classification difficulty limits its applicability to only a few datasets.

In this work, we argue that relying solely on the accuracy criterion can be misleading for evaluating algorithms, even when data complexity measures are considered. We define a model as **invalid** if its error rate is *lower* than the Bayes error rate (BER) or if its accuracy exceeds the Golden Accuracy (GSA = 1 - BER). We propose an algorithm for model *validation* that utilizes BER to assess the effectiveness and reliability of classification models.

This paper is divided into the following sections. We address BER in Section 2. Misleading in evaluation of classifiers is expressed in Section 3. An algorithm for model validation is proposed in Section 4. The work is concluded in Section 5.

2 Bayes error rate

The performance of the best classifier is defined as GSA. For an evaluation to be considered valid, the error rates of classifiers should be greater than or equal to BER. However, in practice, some classifiers may exhibit error rates below BER, which is theoretically impossible and thus undesirable. We recall that Bayes classifier error equals BER and can be calculated in the Gaussian case.

2.1 Bayes classifier

A classification task includes $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d)^\intercal$ as a feature vector on a d-dimensional space $\mathscr{X} = \mathbb{R}^d$, which is labeled over binary random variable Y [16]. A binary classification classifier is a function

$h : \mathscr{X} \to Y \in \mathscr{Y} = \{0,1\}$. Classification error (CE) occurs when an observation of $\boldsymbol{X}$, i.e., $\boldsymbol{x}$, is not correctly assigned to its true class $\mathrm{CE}_h = \Pr(h(\boldsymbol{X}) \neq Y)$ [9]. According to the Bayes' theorem, we recall the following formula:

$$\Pr(Y = y \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{\Pr(Y = y) f_{\boldsymbol{X}|Y}(\boldsymbol{x} \mid y)}{f_{\boldsymbol{X}}(\boldsymbol{x})}, \tag{1}$$

where $f_{\boldsymbol{X}|Y}(\boldsymbol{x}|y)$ is the class-conditional distributions $\Pr(Y = 0) = \pi_0$ and $\Pr(Y = 1) = \pi_1 = 1 - \pi_0$ are prior label probabilities and we have

$$\Pr(Y = 0 \mid \boldsymbol{x}) > \Pr(Y = 1 \mid \boldsymbol{x}) \Longleftrightarrow \pi_0 f(\boldsymbol{x} \mid Y = 0) > \pi_1 f(\boldsymbol{x} \mid Y = 1), \tag{2}$$

[4]. Based on the logical classification strategy, we assign an observation $\boldsymbol{x}$ to the first class with a greater posterior probability [17], and using (1) and (2), Bayes classifier h_B is defined as follows

$$h_B = \begin{cases} 0 & \pi_0 f_{\boldsymbol{X}|Y}(\boldsymbol{x} \mid Y = 0) \geq \pi_1 f_{\boldsymbol{X}|Y}(\boldsymbol{x} \mid Y = 1) \\ 1 & \text{otherwise} \end{cases}. \tag{3}$$

Lemma 1. ([17]) *The Bayes classifier (3) is optimal. Specifically, for any other classification rule h, the classification error of the Bayes classifier satisfies* $CE_{h_B} \leq CE_h$.

2.2 Discriminant function

According to (1), the discriminant function is defined as

$$d_y(\boldsymbol{x}) = \log f_{\boldsymbol{X}|Y}(\boldsymbol{x} \mid y) + \log \Pr(Y = y), \tag{4}$$

and quadratic discriminant g is defined by

$$g(\boldsymbol{x}) = \boldsymbol{x}^\intercal \mathbf{A} \boldsymbol{x} + \boldsymbol{a}^\intercal \boldsymbol{x} + c, \tag{5}$$

where $\mathbf{A}$ is a $d \times d$ matrix, $\boldsymbol{a}$ is a column vector with length d, and c is a constant [3].
Gaussian Classifiers: Let the conditional density of classes be Gaussian as follows

$$f_{\boldsymbol{X}|Y}(\boldsymbol{x} \mid y) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_y|^{\frac{1}{2}}} \exp\Big(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_y)^\intercal \boldsymbol{\Sigma}_y^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_y)\Big), \tag{6}$$

where $\boldsymbol{\Sigma}_y$ and $\boldsymbol{\mu}_y$ are the covariance matrix and mean vector, respectively. By replacing (6) in (4) and simplifying, we have

$$d_y(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_y)^\intercal \boldsymbol{\Sigma}_y^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_y) - \frac{1}{2}\log|\Sigma_y| + \log \Pr(Y = y),$$

[11]. Suppose that the parameters of Gaussian distributions in (6) are known. Using (5), the Bayes classifier in (3) is reduced to

$$g_{\mathrm{B}}(\boldsymbol{x}) = \boldsymbol{x}^\intercal \mathbf{A}_{\mathrm{B}} \boldsymbol{x} + \boldsymbol{a}_{\mathrm{B}}^\intercal \boldsymbol{x} + c_{\mathrm{B}}, \tag{7}$$

where

$$\mathbf{A}_{\mathrm{B}} = -\frac{1}{2}\left(\Sigma_1^{-1} - \Sigma_0^{-1}\right), \ \mathbf{a}_{\mathrm{B}} = \Sigma_1^{-1}\boldsymbol{\mu}_1 - \Sigma_0^{-1}\boldsymbol{\mu}_0,$$

and

$$c_{\mathrm{B}} = -\frac{1}{2}\left(\boldsymbol{\mu}_1^T\Sigma_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^T\Sigma_0^{-1}\boldsymbol{\mu}_0\right) + \frac{1}{2}\log\left(\frac{|\Sigma_0|}{|\Sigma_1|}\right) + \log\left(\frac{1-\pi_0}{\pi_0}\right),$$

[6]. Discriminant function (7) is known as quadratic discriminant analysis (QDA).

Lemma 2. *Let data have multivariate Gaussian distribution with known parameters. The error of* QDA in (7) equals BER for Gaussian models.

Proof. From Lemma 1, we have $\Pr(h_B(\boldsymbol{X}) \neq Y) = \mathrm{BER}$, and the proof is complete. □

Example 1. (Bivariate Gaussian). Assume the class density function (6) has the following parameters:

$$\boldsymbol{\mu}_0 = \begin{bmatrix} -0.2 \\ -0.2 \end{bmatrix}, \boldsymbol{\mu}_1 = \begin{bmatrix} 0.2 \\ 0.2 \end{bmatrix}, \text{ and } \Sigma_0 = \Sigma_1 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}.$$

Consider $\pi_0 = \pi_1$ and the region R as follows

$$R = \{\boldsymbol{x} | \pi_0 f_{\boldsymbol{X}|Y}(\boldsymbol{x} \mid Y = 0) \geq \pi_1 f_{\boldsymbol{X}|Y}(\boldsymbol{x} \mid Y = 1)\}, \tag{8}$$

then

$$\mathrm{BER} = \Pr(h_B(\boldsymbol{X}) \neq Y) = \int_{R^c} \pi_0 f_{\boldsymbol{X}|Y}(\boldsymbol{x} \mid Y = 0)\, d\boldsymbol{x} + \int_R \pi_1 f_{\boldsymbol{X}|Y}(\boldsymbol{x} \mid Y = 1)\, d\boldsymbol{x} = 0.414131$$

Example 2. The location parameters in Example 1 were shifted as follows

$$\boldsymbol{\mu}_0 = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \text{ and } \boldsymbol{\mu}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

and so BER is reduced to 0.139038.

Classifier (7) attains the BER. It is defined as the best theoretical model (BTM) classifying data generated from the Gaussian model. In general, when the class distribution is Gaussian with known parameters, the QDA error equals BER (Lemma 2).

Example 3. Let $f(\boldsymbol{x}|\underset{y=0,1}{\boldsymbol{y}} = y) \sim Cauchy(\boldsymbol{\theta}_y, \boldsymbol{\zeta}_y)$, be class conditional density function as follows:

$$f(\boldsymbol{x}|\boldsymbol{y} = y) = \frac{\Gamma[(1+k)/2]}{\Gamma(1/2)\pi^{k/2}|\zeta_y|^{1/2}\left[1 + (\boldsymbol{x} - \boldsymbol{\theta}_y)^T \zeta_y^{-1}(\boldsymbol{x} - \boldsymbol{\theta}_y)\right]^{(1+k)/2}},$$

where $\boldsymbol{\theta}_y$ is a location vector, $\boldsymbol{\zeta}_y$ is a $k \times k$ positive-definite dispersion matrix, and $\pi = 3.14$ is a constant. Assume $k = 2$,

$$\boldsymbol{\theta}_0 = \begin{bmatrix} -0.1 \\ -0.1 \end{bmatrix}, \boldsymbol{\theta}_1 = \begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix}, \text{ and } \boldsymbol{\zeta}_0 = \boldsymbol{\zeta}_1 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix},$$

then by considering R in (8)

$$\mathrm{BER} = \Pr(h_B(\boldsymbol{X}) \neq Y) = \int_{R^c} \pi_0 f_{\boldsymbol{X}|Y}(\boldsymbol{x} \mid Y = 0)\, d\boldsymbol{x} + \int_R \pi_1 f_{\boldsymbol{X}|Y}(\boldsymbol{x} \mid Y = 1)\, d\boldsymbol{x} = 0.465609.$$

3 Why accuracy-only evaluation can be misleading

An evaluation is not misleading when the accuracy of classifiers is less than the accuracy of BTM. To show the misleading of classifiers evaluation, we generate datasets from the Gaussian model in Example 1 with the same complexity (density, separability, and balanced weights). Figure 1 visualizes the contour and scatter plot of the model and a data sample.

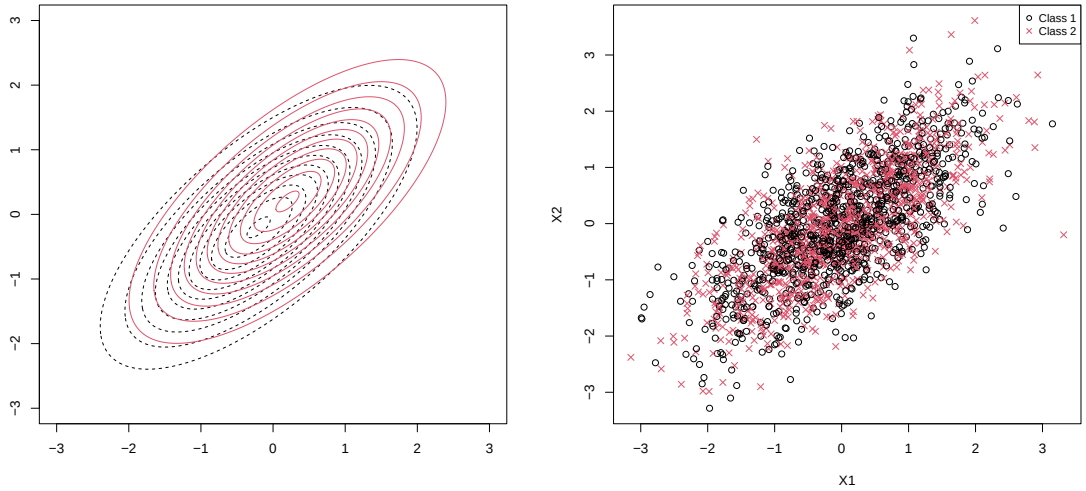

Figure 1: The Left graphs show contour plots of bivariate Gaussian distributions and the right graph is a scatter plot of generated data in Example 1.

To analyze the performance of the classifiers, we consider QDA, k-Nearest Neighbors (kNN), Logistic Regression (LR), Support Vector Machine for Classification (SVC), Gaussian Naive Bayes (GNB), Linear classifiers with Stochastic Gradient Descent (SGD) training, Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), Ada Boost (AB), Multi-layer Perceptron (MLP), algorithms from Scikit-learn [14] and parameters were taken as the defaults. Accuracy for the classification tasks is computed using 10-fold Cross-Validation. A classifier from among these 11 algorithms that demonstrates the minimum error on a dataset is designated as the Best Empirical Method (BEM).

In Figures 2 and 3, we compare the efficiency of classification algorithms with BTM represented by a solid blue line and BER depicted by a blue dashed line. As the sample size increases, the error rate of BEM, indicated by the red line, tends to converge towards the BTM. In most cases, the error of BEM is lower than both BTM and BER. However, these scenarios lead to potentially misleading evaluations of classifier performance. A classifier is often selected as superior because its error is the least. If the error is less than BER, this outcome is theoretically impossible. This discrepancy suggests a misjudgment in assessing classifier efficiency and necessitates a careful review of evaluation metrics and methodologies.

We summarize the results of Figure 3 in Table 1, in which BER equals 0.139038. In samples greater than 300, all algorithms are valid (the accuracy of classifiers is not less than BER); however, we are misled in evaluating all datasets except 500 and 10000.

Focusing on Figure 4 (top left graph, $N = 300$), a detail of the top left graph in Figure 2, one can see that MLP algorithm error is BEM and less than BTM error that leads to decision-making errors. Top right graph ($N = 500$), the SVC algorithm error is BEM and less than BER, which means that,

at the same time, we make a mistake in making a decision (misleading) and have an invalid model. The results are summarized in Table 2. According to the top left graph in Figure 2, Figure 4, and BER criterion with the same data complexity, we have misleading in any sample sizes. However, by increasing sample size, BEM error tends to BTM error.

Figure 5 shows the performance of classifiers on the dataset generated from a bivariate Cauchy distribution with the parameters in Example 3. MLP and GB are invalid models and potentially can be misleading.

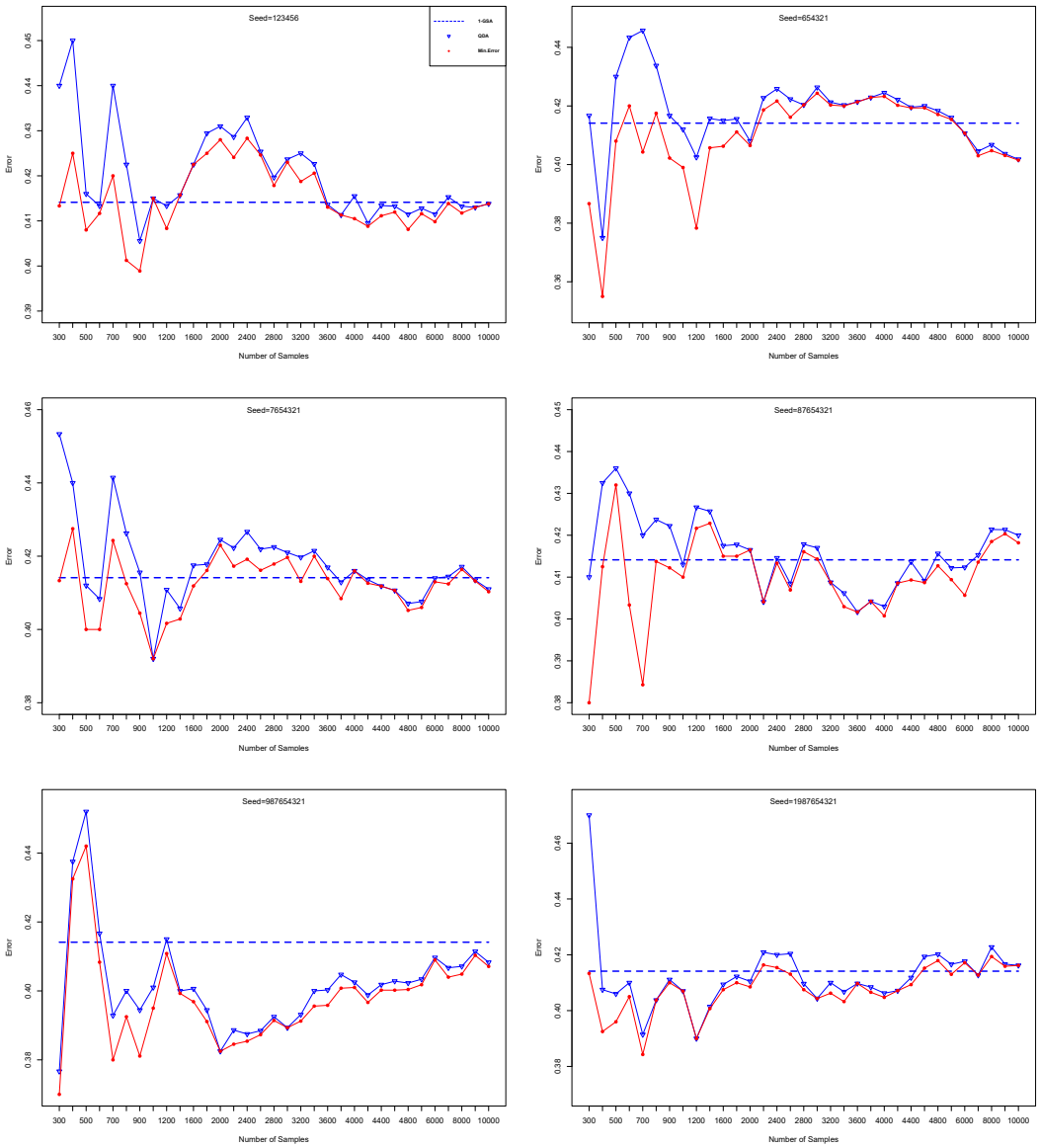

Figure 2: Comparison of classification errors: BEM error is depicted by a red line with solid circles, BER by a blue dashed line, and BTM error by a blue line with inverted triangles. Data samples were generated from Example 1 using different seeds but maintained consistent complexity. Errors falling below the BER line are paradoxical. Instances where BEM error is less than BTM error can lead to misleading and incorrect decision.

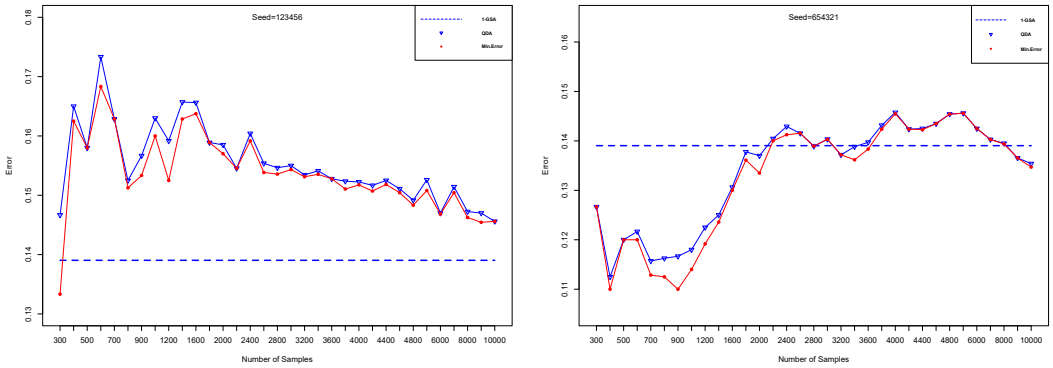

Figure 3: Comparison of classification errors: BEM error is depicted by a red line with solid circles, BER by a blue dashed line, and BTM error by a blue line with inverted triangles. Data samples were generated from Example 2 using different seeds but maintained consistent complexity. Errors falling below the BER line are paradoxical. Instances where BEM error is less than BTM error can lead to misleading and wrong decisions.

Table 1: Results summary of the left graph of Figure 3. The misleading cases and invalid models are highlighted in red for each dataset.

Sample Size	Best Empirical Method	Underestimate w.r.t. BER	Valid	Underestimate w.r.t. Best Theoretical Method	Misleading
300	MLP	Yes	No	Yes	Yes
500	QDA	No	Yes	No	No
800	LR	No	Yes	Yes	Yes
1200	SVC	No	Yes	Yes	Yes
2400	LR	No	Yes	Yes	Yes
4000	SVC	No	Yes	Yes	Yes
7000	MLP	No	Yes	Yes	Yes
10000	QDA	No	Yes	No	No

Table 2: Summary of graph results as shown in Figure 4. Cases and models that did not meet validation criteria are highlighted in red.

Sample Size	Best Empirical Method	Underestimate w.r.t. BER	Valid	Underestimate w.r.t. Best w.r.t. Best Theoretical Method	Misleading Misleading
300	MLP	Yes	No	Yes	Yes
500	SVC	Yes	No	Yes	Yes
800	SVC	Yes	No	Yes	Yes
1200	LR	Yes	No	Yes	Yes
2400	LR	No	Yes	Yes	Yes
4000	GNB	Yes	No	Yes	Yes
7000	LR	Yes	No	Yes	Yes
10000	LR	Yes	No	Yes	Yes

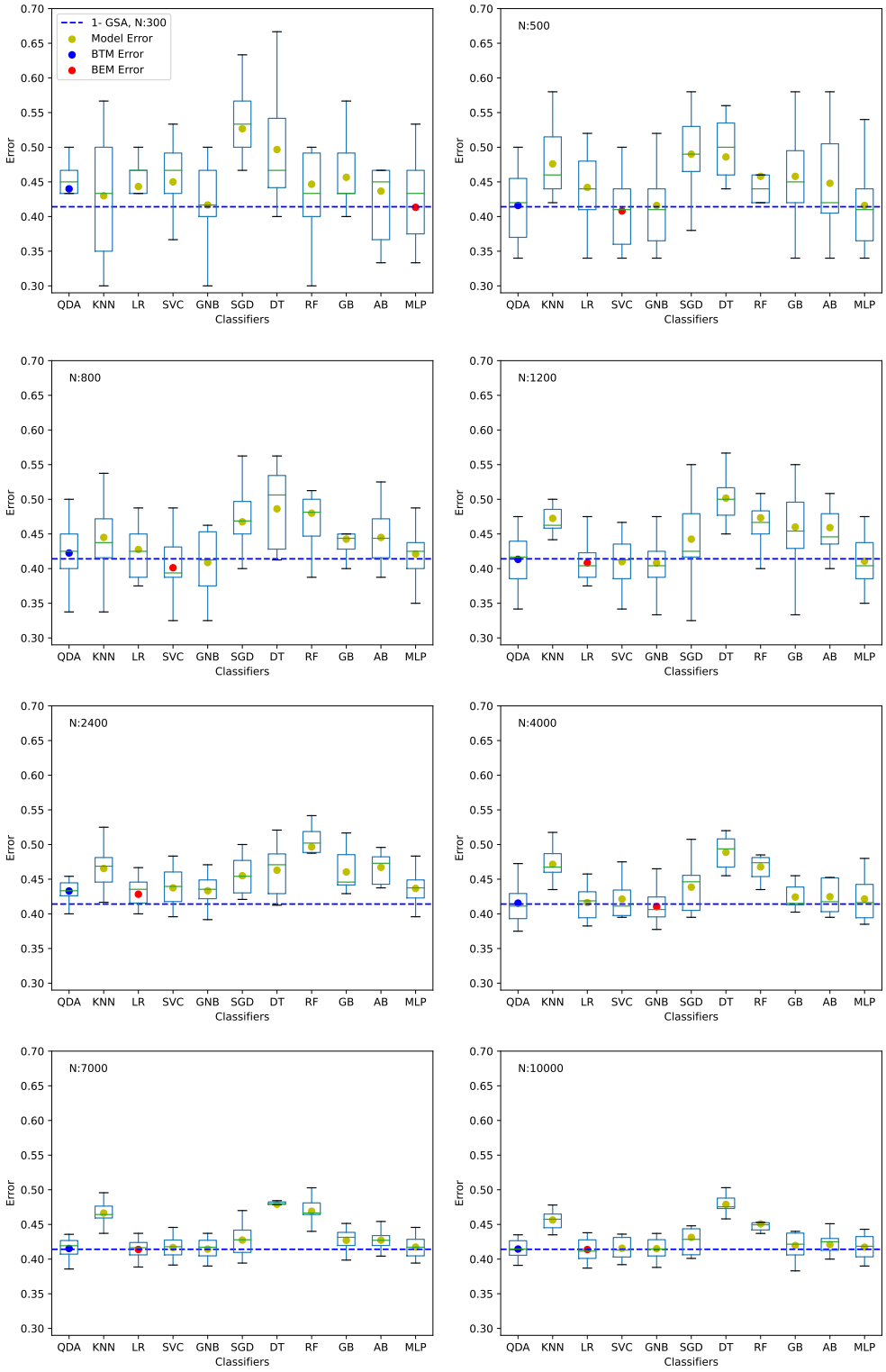

Figure 4: Performance of classifiers based on different samples from the top left graph in Figure 2. BTM and BEM errors are depicted in blue and red, respectively. A blue dashed line represents the BER. The models with the error under the BER line are paradoxical. Classifiers whose error is less than BTM are where we go wrong.

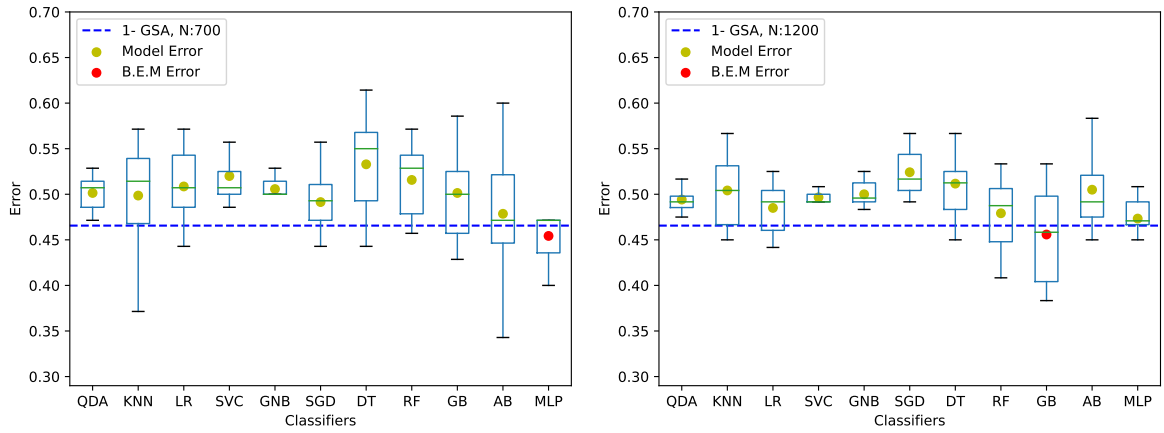

Figure 5: Evaluating performance of classifiers on generated data from bivariate Cauchy distributions in Example 3. The models with the error under the BER line are paradoxical.

4 Model validation

The validity of the models is determined by comparing them with BER. Models with errors less than BER are considered invalid. The estimated BER ($\hat{\text{BER}}$) equals 0.00 ± 0.01 for the Iris dataset so that all algorithms will be valid. The validation process is done through Algorithm 1.

Algorithm 1 Model Validation

1: Compute $\hat{\text{BER}}$ for a given dataset.
2: Compute the classification error of the candidate classifier.
3: If the classification error in step 2 is less than $\hat{\text{BER}}$, the classification is invalid.
4: Repeat step 2 to find a list of valid models.
5: Select a classifier with minimum classification error in step 4.

Remark. To accurately estimate BER and its margin of error, one can employ the advanced methodology developed by Noshad et al. [13].

5 Discussion

In this note, we show that by considering the same complexity measures on datasets, the accuracy criterion derived from the confusion matrix for evaluating classifiers is misleading. In small samples, the evaluation of models leads to overestimation. The misleadingness of this approach in large samples still exists; however, the strength is that the error of BEM tends to the error of BTM. Another unacceptable observation is that the accuracy of the classification algorithms is paradoxically less than BER, which is theoretically impossible and leads to invalid evaluation. To solve this issue, it is necessary to compare the error of algorithms with the BER. However, it is impossible to calculate or compute this value in high-dimensional cases. Therefore, we propose that the evaluation and validation of classification algorithms should include comparisons with the BER to ensure accurate, non-misleading results.

Acknowledgment. We thank the associate editor and referees for their rigorous reviews and valuable insights, which have greatly enhanced the manuscript's quality and accuracy.

References

[1] E.Alpaydin, *Introduction to Machine Learning*, Adaptive Computation and Machine Learning, MIT Press, Third Edition, 2014.

[2] C.M. Bishop, *Pattern Recognition and Machine Learning*, Volume 4 of Information Science and Statistics, Springer, 2006.

[3] L. Dalton, E. Dougherty, *Optimal Bayesian Classification*, Press Monograph Series, SPIE Press, 2020.

[4] R. Duda, P. Hart, D. Stork, *Pattern Classification*, Wiley, 2012.

[5] A. Fernández, S. García, M. Galar, R.C. Prati, B. Krawczyk, F. Herrera, *Data Intrinsic Characteristics*, pages 253–277, Springer, 2018.

[6] K. Fukunaga, *Introduction to Statistical Pattern Recognition, Chapter 10*, Academic Press, 1990.

[7] S. Guan, M.H. Loew, *A novel intrinsic measure of data separability*, **52** (2022) 17734–17750.

[8] T.K. Ho, M. Basu, *Complexity measures of supervised classification problems,* IEEE Trans. Pattern Anal. Mach. Intell. **24** (2002) 289–300.

[9] A. Izenman, *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*, Springer Texts in Statistics. Springer, 2009.

[10] A.C. Lorena, L.P.F. Garcia, J. Lehmann, M.C.P. Souto, T.K. Ho, *How complex is your classification problem? A survey on measuring classification complexity,* ACM Comput. Surv. **52** (2019) 1–34.

[11] G.J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, 2004.

[12] K. Murphy, *Probabilistic Machine Learning: An Introduction*, Adaptive Computation and Machine Learning series, MIT Press, 2022.

[13] M. Noshad, L. Xu, A. Hero, *Learning to benchmark: Determining best achievable misclassification error from training data*, 2019, arXiv:1909.07192.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *Scikit-learn: Machine learning in Python.* J. Mach. Learn. Res. **12** (2011) 2825–2830.

[15] Y. Peleg, *Hungabunga: Brute-Force all sklearn models with all possible hyperparameters, and rank using cross-validation*, GitHub, Retrieved from https://github.com/ypeleg/HungaBunga, 2023.

[16] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*, Elsevier, 2020.

[17] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*, Springer, 2004.

[18] L. Xue, X. Zhang, W. Jiang, K. Huo, Q. Shen, *A classification performance evaluation measure considering data separability* In L. Iliadis, A. Papaleonidas, P. Angelov, and C. Jayne, editors, Artificial Neural Networks and Machine Learning – ICANN 2023, pages 1–13, Springer Nature Switzerland, 2023.

[19] S. Yu, X. Li, Y. Feng, X. Zhang, S. Chen. *An instance-oriented performance measure for classification.* Inf. Sci. **580** (2021) 598–619.