JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 00, NO. 0, 1–11: Theory and Methods
https://doi.org/10.1080/01621459.2021.1979010

Is a Classification Procedure Good Enough?—A Goodness-of-Fit Assessment Tool for Classification Learning

Jiawei Zhang, Jie Ding, and Yuhong Yang

School of Statistics, University of Minnesota, Minneapolis, MN

ABSTRACT

In recent years, many nontraditional classification methods, such as random forest, boosting, and neural network, have been widely used in applications. Their performance is typically measured in terms of classification accuracy. While the classification error rate and the like are important, they do not address a fundamental question: Is the classification method underfitted? To our best knowledge, there is no existing method that can assess the goodness of fit of a general classification procedure. Indeed, the lack of a parametric assumption makes it challenging to construct proper tests. To overcome this difficulty, we propose a methodology called BAGofT that splits the data into a training set and a validation set. First, the classification procedure to assess is applied to the training set, which is also used to adaptively find a data grouping that reveals the most severe regions of underfitting. Then, based on this grouping, we calculate a test statistic by comparing the estimated success probabilities and the actual observed responses from the validation set. The data splitting guarantees that the size of the test is controlled under the null hypothesis, and the power of the test goes to one as the sample size increases under the alternative hypothesis. For testing parametric classification models, the BAGofT has a broader scope than the existing methods since it is not restricted to specific parametric models (e.g., logistic regression). Extensive simulation studies show the utility of the BAGofT when assessing general classification procedures and its strengths over some existing methods when testing parametric classification models. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received January 2021
Accepted September 2021

KEYWORDS
Adaptive partition;
Classification procedure;
Goodness-of-fit test

1. Introduction

The development of various classification procedures has been a backbone of the contemporary learning toolbox to solve various data challenges. This work addresses the following fundamental problem in classification learning: *How to assess whether a classification procedure is good enough, in the sense that it has no systematic defects, as reflected in its convergence to the data-generating process, for given data?*

We highlight that the assessment raised in the above question is fundamentally different from assessing the predictive performance. In most applications, a classification procedure's performance is often assessed based on its *classification accuracy* on preset validation data or through cross-validation. Conceptually, the predictive accuracy does not characterize a procedure's deviation from the underlying data generating process per se. For instance, when the conditional probability function of success given the covariates is simply 0.5, the best possible classifier is a random guess, which provides a low classification accuracy.

It is critical to address the above question in several emerging learning scenarios where the classification accuracy alone cannot solve the problems. For example, an increasing number of entities use Machine-Learning-as-a-Service (MLaaS) (Ribeiro, Grolinger, and Capretz 2015) or cooperative learning protocols (Xian et al. 2020) to train a model from paid cloud-computing services. It is economically significant to decide whether the current learning method has a significant discrepancy from the data and needs to be further improved. Another example concerns the use of "benchmark data" for comparing classification procedures, for example, those from Kaggle (*https://rb.gy/bvepug*) or UCI (*https://archive.ics.uci.edu/ml/datasets.php*). Based on the validation accuracy as an evaluation metric, the winning procedure selected from many candidate learners may have already been overfitting luckily and deviating from the underlying data generating process. In this case, assessing the deviation of the learning procedures from the data distribution is also very helpful.

In dealing with a parametric model, the existing literature addresses the above problem from a goodness-of-fit (GOF) test perspective. For binary regression, two classical approaches are the Pearson's chi-squares (χ^2) test and the residual deviance test, which group the observations according to distinct covariate values. When the number of observations in each group is small, for example, there is at least one continuous covariate, the above two tests cannot be applied. Various tests have been developed to address this issue. These include the tests based on the distribution of the Pearson's χ^2 statistic under sparse data (McCullagh 1985; Osius and Rojek 1992; Farrington 1996), kernel smoothed residuals (Le Cessie and Van Houwelingen 1991), the comparison with a generalized model (Stukel 1988), the comparison between an estimator from the control data and an estimator from the joint data in the context of case-control studies (Bondell 2007), the Pearson-type statistics calculated

CONTACT Yuhong Yang *yyang@stat.umn.edu* School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church St. SE, Minneapolis, MN 55455.
This article has been corrected with minor changes. These changes do not impact the academic content of the article.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

from bootstrap samples (Yin and Ma 2013), information matrix tests (White 1982; Orme 1988), the grouping of observations into a finite number of sets (Hosmer and Lemeshow 1980; Pigeon and Heyse 1999; Pulkstenis and Robinson 2002; Xie, Pendergast, and Clarke 2008; Liu, Nelson, and Yang 2012), and the predictive log-likelihood on validation data (Lu and Yang 2019).

However, there are two weaknesses of the existing GOF tests for parametric classification models. First, most tests only control the Type I error probability, but without theoretical guarantees on the test power. Second, the existing methods focus on the GOF of specific models, such as logistic regression, and may not be applied to general binary regression models.

For general classification procedures such as decision trees, neural networks, k-nearest neighbors, and support vector machines, to our best knowledge, there is no existing method to assess their GOF. We broaden the notion of the GOF test to address the question above for general classification procedures.

We propose a new methodology named the binary adaptive goodness-of-fit test (BAGofT) for testing the GOF of both parametric classification models and general classification procedures. The developed tools may guide data analysts to understand whether a given procedure, possibly selected from a set of candidates, deviates significantly from the underlying data distribution. We focus on assessing binary classification procedures that provide estimates of the conditional probability function.

The BAGofT employs a data splitting technique, which helps the test overcome the difficulties in the general setting where there is no workable saturated model to compare with, as used in Pearson's chi-squares and deviance-based tests. On the 'training' set, the BAGofT applies an adaptive partition of the covariate space that highlights the potential underfitting of the model or procedure to assess. Then, the BAGofT calculates a Pearson-type test statistic on the remaining 'validation' part of the data based on the grouping from the adaptive partition.

For parametric classifications, the BAGofT enjoys theoretical guarantees for its consistency under a broad range of alternative hypotheses, including those concerning misspecified covariates and model structures. Its adaptive partition can flexibly expose different kinds of weaknesses from the parametric classification model to test. Importantly, unlike the previous methods, it allows the number of groups to grow with the sample size when a finer partition is needed. Moreover, the probability of the Type I error is well controlled due to data splitting.

For a general classification procedure without a workable benchmark to compare with, one major challenge is to define the GOF. Unlike parametric models, whose convergence is well understood, general classification procedures can have different convergence rates. If we choose the splitting ratio of the BAGofT according to a specific rate, then the size of the test can be controlled as long as the procedure to assess converges not slower than the specified rate under the null hypothesis; the BAGofT consistently rejects the hypothesis otherwise. In practice, since the convergence rate of the procedure to assess is unknown, we advocate a method based on the BAGofT with multiple data splitting ratios. Our experimental results show that this method can faithfully reveal possible moderate or severe deficiency of a classification procedure.

The outline of the article is given as follows. In Section 2, we provide the background of the problem. In Section 3, we introduce the BAGofT and establish its properties. In Section 4, we provide some practical guidelines on implementing the BAGofT. We present simulation results in Section 5 and real data examples in Section 6. We conclude the paper in Section 7. The proofs and additional numerical results are included in the supplementary material.

2. Problem Formulation

2.1. Setup

Let Y be the binary response variable that takes 0 or 1, and $\boldsymbol{X}$ be the vector of p covariates. The support of $\boldsymbol{X}$ is $\mathbb{S} \subseteq \mathbb{R}^p$. Let $\pi(\cdot)$ be the conditional probability function:

$$\pi(\boldsymbol{x}) = P(Y = 1|\boldsymbol{X} = \boldsymbol{x}),\ \boldsymbol{x} \in \mathbb{S}. \tag{1}$$

The data, denoted by D_n, consist of n iid observations from a population distribution of the pair $(Y, \boldsymbol{X})$. The conditional probability function $\pi(\cdot)$ is allowed to change with the sample size. We denote the fitted conditional probability function obtained from a classification model (or a procedure) on D_n by $\hat{\pi}_{D_n}(\cdot)$.

2.2. Testing Parametric Classification Models

A parametric classification model assumes that $\pi(\cdot) = f(\cdot, \boldsymbol{\beta})$, where f is known and the unknown parameter $\boldsymbol{\beta}$ is in a finite-dimensional set $\mathbb{B}$. For example, a generalized linear model assumes that $f(\boldsymbol{x}, \boldsymbol{\beta}) = g^{-1}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})$, where $g(\cdot)$ is a link function. The null and alternative hypotheses of the GOF for testing a parametric classification model are defined by

$$H_0:\ \pi(\cdot) \in \{f(\cdot, \boldsymbol{\beta}) \mid \boldsymbol{\beta} \in \mathbb{B}\}, \quad H_1:\ \pi(\cdot) \notin \{f(\cdot, \boldsymbol{\beta}) \mid \boldsymbol{\beta} \in \mathbb{B}\}.$$

We refer to the parametric classification model to assess as MTA.

2.3. Assessing General Classification Procedures

Compared with parametric classification models, general classification procedures are not restricted to be in a parametric form. They include any modeling technique that maps the data D_n to a fitted conditional probability function $\hat{\pi}_{D_n}(\cdot) : \mathbb{S} \to [0, 1]$. For a general classification procedure, the convergence rate of $\hat{\pi}_{D_n}(\cdot)$ is essential from a theoretical viewpoint. Let r_n be the convergence rate of the classification procedure we assess under the null hypothesis. The null and alternative hypotheses of the GOF test for a general classification procedure to assess (PTA) are

$$H_0 : \sup_{\boldsymbol{x} \in \mathbb{S}} |\hat{\pi}_{D_n}(\boldsymbol{x}) - \pi(\boldsymbol{x})| = O_p(r_n),$$

$$H_1 : \exists\ \mathbb{M}_n \subseteq \mathbb{S} \text{ with } P(\boldsymbol{x} \in \mathbb{M}_n) \text{ bounded away from 0 such that } \inf_{\boldsymbol{x} \in \mathbb{M}_n} |\hat{\pi}_{D_n}(\boldsymbol{x}) - \pi(\boldsymbol{x})|/r_n \to_p \infty,$$

as $n \to \infty$, where the set $\mathbb{M}_n$ may change with n. So under H_0, $\hat{\pi}_{D_n}(\cdot)$ converges to $\pi(\cdot)$ not slower than r_n, and under H_1, it converges slower (or does not converge) to $\pi(\cdot)$.

3. Binary Adaptive Goodness-of-fit Test (BAGofT)

3.1. Test Statistic

The BAGofT is a two-stage approach where the first stage explores a data-adaptive grouping and the second stage performs testing based on that grouping. The adaptive grouping consists of the following steps. 1. Split the data into a training set D_{n_1} with size n_1 and a validation set D_{n_2} with size n_2. 2. Apply the MTA or PTA to D_{n_1} and obtain the estimated probabilities for both the training set and validation set. 3. Generate a partition $\{\hat{G}_{D_{n_1},1}, \ldots \hat{G}_{D_{n_1},K_n}\}$ of the support $\mathbb{S}$. This partition can be obtained by any method that meets the following two requirements. (i) Denote the set of responses and covariates in D_{n_2} by D_{y_e} and D_{x_e}, respectively. The partition needs to be independent of D_{y_e} conditional on D_{x_e}. It means that we may obtain a partition based on the performance of the MTA/PTA on the training set. We can also use D_{x_e} to control the group sizes for the partition of D_{n_2}. (ii) The number of groups $K_n \geq 2$. Note that K_n can be data-driven and is not required to be uniformly upper bounded. We propose an adaptive partition algorithm, which is elaborated in Section 4.2. 4. Group D_{n_2} into sets based on the obtained partition. Let $\boldsymbol{x}_{e,i}$ $(i = 1, \ldots, n_2)$ denote the covariates observations from the validation set. For $i = 1, \ldots, n_2$, the ith observation in the validation set is said to belong to group k, if $\boldsymbol{x}_{e,i} \in \hat{G}_{D_{n_1},k}$.

For the testing stage, let $R_i = y_{e,i} - \hat{\pi}_{D_{n_1}}(\boldsymbol{x}_{e,i})$, $\sigma_i^2 = \hat{\pi}_{D_{n_1}}(\boldsymbol{x}_{e,i})\left\{1 - \hat{\pi}_{D_{n_1}}(\boldsymbol{x}_{e,i})\right\}$, where $i = 1, \ldots, n_2$ and $y_{e,i}$ is the response observation from the validation set, and

$$T = \sum_{k=1}^{K_n} \left(\frac{\sum_{\{i:\, \boldsymbol{x}_{e,i} \in \hat{G}_{D_{n_1},k}\}} R_i}{\sqrt{\sum_{\{i:\, \boldsymbol{x}_{e,i} \in \hat{G}_{D_{n_1},k}\}} \sigma_i^2}} \right)^2.$$

We define the following p-value statistic based on the CDF of the chi-squared distribution with degrees of freedom K_n:

$$\text{BAG} = 1 - P(\chi^2_{K_n} \leq T | T, K_n). \tag{2}$$

We reject H_0 when BAG is less than the specified significance level, since BAG tends to be small when the discrepancy between $\hat{\pi}_{D_{n_1}}(\cdot)$ and $\pi(\cdot)$ as quantified by T is large.

Compared with the Hosmer–Lemeshow test and other relevant methods, the proposed method allows desirable features such as pre-screening candidate grouping methods (we do not need the Bonferroni correction when considering different groupings), incorporating prior or practical knowledge that is potentially adversarial to the MTA or PTA (e.g., the BAGofT partition can be based on some potentially important variables not in the MTA or PTA), and providing interpretations on the data regions where the MTA or PTA is likely to fail. The above flexibility often leads to a significantly improved statistical power (elaborated in Section 4.2). It is worth noting that the BAGofT exhibits a tradeoff in data splitting. On the one hand, sufficient validation data used to perform tests can enhance power due to a more reliable assessment of the deviation. On the other hand, more training data enables a better estimation of $\pi(\cdot)$ and the selection of an adversarial grouping that increases power. We will develop theoretical analyses and experimental studies to guide the use of an appropriate splitting ratio.

3.2. Theory for Testing Parametric Classification Models

We first establish a theorem that the BAGofT p-value statistic converges in distribution to the standard uniform distribution under H_0, which asymptotically guarantees the size of the test. We need the following technical conditions.

For positive sequences a_n and b_n, we write $a_n = \omega(b_n)$ if $a_n/b_n \to \infty$ as $n \to \infty$.

Condition 1 (*Sufficient number of observations in each group*). There exists a positive sequence $\{\underline{m}_n\}$ such that $\min_{k=1,\ldots,K_n} \sum_{i=1}^{n_2} I\{\boldsymbol{x}_{e,i} \in \hat{G}_{D_{n_1},k}\} \geq \underline{m}_n$ *a.s.*, and $\underline{m}_n = \omega({n_2}^{5/7})$ as $n \to \infty$.

Condition 2 (*Bounded true probabilities*). There exists a positive constant $0 < c_1 < 1/2$ such that $c_1 \leq \pi(\boldsymbol{x}) \leq 1 - c_1$ for all $\boldsymbol{x} \in \mathbb{S}$.

Condition 3 (*Parametric rate of convergence under H_0*). Under H_0, $\sup_{\boldsymbol{x} \in \mathbb{S}} |\hat{\pi}_{D_n}(\boldsymbol{x}) - \pi(\boldsymbol{x})| = O_p\left(1/\sqrt{n}\right)$ as $n \to \infty$.

Condition 1 is mild and can be guaranteed by merging small-sized groups on D_{n_2}. Condition 2 is a technical requirement so that the Pearson residuals in the theoretical derivations would be bounded, which is satisfied, for example, under the GLM framework with compact parameter and covariates spaces, and it can be relaxed if more assumptions are made on the tail of the covariate distributions. Condition 3 holds for a typical parametric model and a compact set $\mathbb{S}$.

Throughout the article, we let U denote the standard uniform distribution.

Theorem 1 (*Convergence of* BAG *for parametric models under H_0*). Assume that Conditions 1–3 hold. Under H_0, if $n_1, n_2 \to \infty$ and $n_2 = o(n_1^{3/5})$ as $n \to \infty$, we have $\text{BAG} \to_d U$.

Accordingly, if we reject the MTA when the BAGofT p-value statistic is less than 0.05, we obtain the asymptotic size 0.05. The requirement of n_1 and n_2 in the above theorem indicates that the number of observations for estimating the parameters and forming groups (n_1) needs to be much larger than the number for performing tests (n_2). Otherwise, the deviation introduced by a random fluctuation due to a small training size (instead of true misspecification) may be picked up by the BAGofT. It is interesting to note that this data splitting ratio direction is opposite to that for the consistent selection of the best classification procedure via cross-validation (Yang 2006; see also Yu and Feng 2014), although other splitting ratios in between have been recommended for the purpose of tuning parameter or model selection (Bondell, Krishna, and Ghosh 2010; Lei 2020).

Next, we establish the theorem that shows the BAGofT asymptotically rejects an underfitted model under H_1.

Condition 4 (*Convergence under H_1*). There exists a function $\pi_a : \mathbb{S} \to [0, 1]$, which is not in $\{f(\cdot, \boldsymbol{\beta}) \mid \boldsymbol{\beta} \in \mathbb{B}\}$ and allowed to change with n, such that under H_1,

$$\sup_{\boldsymbol{x} \in \mathbb{S}} |\hat{\pi}_{D_n}(\boldsymbol{x}) - \pi_a(\boldsymbol{x})| \to_p 0 \text{ as } n \to \infty. \tag{3}$$

Moreover, there exists a constant $0 < c_2 < 1/2$ such that $c_2 \leq \pi_a(\boldsymbol{x}) \leq 1 - c_2$ for $\boldsymbol{x} \in \mathbb{S}$.

Condition 5 (Identifiable difference under H_1). Under H_1, with probability going to one, there exists $\mathbb{M}_n \subseteq \mathbb{S}$, which may depend on D_{n_1}, such that

$$\operatorname*{ess\,inf}_{\boldsymbol{x}\in\mathbb{M}_n}(\pi(\boldsymbol{x}) - \pi_a(\boldsymbol{x})) \geq c, \quad \text{or} \tag{4}$$

$$\operatorname*{ess\,sup}_{\boldsymbol{x}\in\mathbb{M}_n}(\pi(\boldsymbol{x}) - \pi_a(\boldsymbol{x})) \leq -c, \tag{5}$$

for a positive constant $c < 1$. We also require that there exists a positive constant $c_0 < c$ such that there is at least one group indexed by k^* with

$$P(\hat{n}_{2,k^*}^{\mathbb{M}_n}/\hat{n}_{2,k^*} > (1+c_0)/(1+c)) \to 1, \tag{6}$$

as $n \to \infty$, where $\hat{n}_{2,k} = \sum_{i=1}^{n_2} I\{\boldsymbol{x}_{e,i} \in \hat{G}_{D_{n_1},k}\}$ denotes the number of validation observations in the kth group, and $\hat{n}_{2,k}^{\mathbb{M}_n} = \sum_{i=1}^{n_2} I\{\boldsymbol{x}_{e,i} \in \hat{G}_{D_{n_1},k} \cap \mathbb{M}_n\}$ denotes the number of validation observations in both the kth group and the set $\mathbb{M}_n$.

Condition 4 requires the convergence of the model under the alternative. We can obtain Equation (3) under the regularity conditions for the convergence of misspecified maximum likelihood estimators (White (1982); specifically for GLM, Fahrmexr (1990)) . Condition 5 guarantees that under H_1, the deviation between the true model and the fitted model can be captured by the adaptive partition. In particular, Equation (6) requires the adaptive selection of a set that contains sufficiently many observations that deviate in the same direction. This is an intuitive and mild condition. The required proportion of observations satisfying Equation (4) or (5) is lowed bounded by $1/(1+c)$, which gets smaller when the bias c is larger. We will provide a practical algorithm in Section 4.2 to adaptively search for the most revealing partition according to the Pearson residual (which measures the discrepancy between $\pi_a(\cdot)$ and $\pi(\cdot)$). Further discussions on how that algorithm meets Condition 5 are included in the supplementary material. Theoretical properties of the algorithm (including the case for assessing general classification procedures) can be found in our discussion section.

Theorem 2 (Consistency of BAG for parametric models under H_1). Suppose that Conditions 1, 2, 4, and 5 hold. Under H_1, if the training and validation sizes satisfy $n_1, n_2 \to \infty$ as $n \to \infty$, we have BAG $\to_p 0$, which implies the consistency of the test.

In applications, we do not know whether H_0 or H_1 holds. If we take n_1 and n_2 such that $n_1, n_2 \to \infty$ and $n_2 = o(n_1^{3/5})$ as $n \to \infty$, under the conditions for Theorems 1 and 2, respectively, then the BAGofT achieves the desired asymptotic Type I error control and consistency in power.

3.3. Theory for Assessing General Classification Procedures

In this section, we establish properties of the BAGofT for general classification procedures.

Condition 6 (Convergence at a general rate under H_0). Under H_0, $\sup_{\boldsymbol{x}\in\mathbb{S}}|\hat{\pi}_{D_n}(\boldsymbol{x}) - \pi(\boldsymbol{x})| = O_p(r_n)$ as $n \to \infty$, with $r_n \to 0$ as $n \to \infty$.

Theorem 3 (Convergence of BAG under H_0 for classification procedures). Under H_0, given Conditions 1, 2, and 6, if $n_2 \to \infty$ and $n_2 = o(r_{n_1}^{-6/5})$ as $n \to \infty$, we have BAG $\to_d U$.

Condition 7 (Existence of an identifiable slow converging set under H_1). Under H_1, with probability going to one, there exists $\mathbb{M}_n \subseteq \mathbb{S}$, which may depend on D_{n_1}, such that $\operatorname{ess\,inf}_{\boldsymbol{x}\in\mathbb{M}_n}(\hat{\pi}_{D_{n_1}}(\boldsymbol{x}) - \pi(\boldsymbol{x})) \geq 0$ or $\operatorname{ess\,sup}_{\boldsymbol{x}\in\mathbb{M}_n}(\hat{\pi}_{D_{n_1}}(\boldsymbol{x}) - \pi(\boldsymbol{x})) \leq 0$, and $\inf_{\boldsymbol{x}\in\mathbb{M}_n}|\hat{\pi}_{D_{n_1}}(\boldsymbol{x}) - \pi(\boldsymbol{x})|/r_{n_1}^{(a)} \geq \zeta$ almost surely, for a positive constant ζ and a positive sequence $r_{n_1}^{(a)} \to 0$ as $n \to \infty$. We also require that there is at least one group indexed by k^* with

$$\frac{\hat{n}_{2,k^*} - \hat{n}_{2,k^*}^{\mathbb{M}_n}}{\hat{n}_{2,k^*} r_{n_1}^{(a)}} \to_p 0, \tag{7}$$

as $n \to \infty$, where $\hat{n}_{2,k}$ and $\hat{n}_{2,k}^{\mathbb{M}_n}$ are defined in Condition 5.

Condition 8 (Bounded predicted probability). There exists a constant $0 < c_3 < 1/2$ such that $c_3 \leq \hat{\pi}_{D_n}(\boldsymbol{x}) \leq 1 - c_3$ almost surely for $\boldsymbol{x} \in \mathbb{S}$ and for all n.

Condition 7 requires the existence of an identifiable region where $\hat{\pi}_{D_n}(\cdot)$ from the PTA converges slowly (or not at all) to the data generating $\pi(\cdot)$ as $n \to \infty$. Further discussions on the theoretical guarantee to identify an $\mathbb{M}_n$ in Condition 7 are included in the supplementary material.

For positive sequences a_n and b_n, we write $a_n = \Omega(b_n)$ if there exists $C > 0$, such that $a_n/b_n \geq C$.

Theorem 4 (Consistency of BAG under H_1 for classification procedures). Under the alternative, assume that Conditions 1, 2, 7, and 8 hold, $n_2 \to \infty$, and $n_2 = \Omega((r_{n_1}^{(a)})^{-6})$ as $n \to \infty$. Then, we have BAG $\to_p 0$ as $n \to \infty$.

The theorem shows that the BAGofT can flag a slow converging classification procedure when there is sufficient validation data. Theorems 3 and 4 imply the following corollary.

Corollary 1 (Obtaining both size control and consistency for learning procedures). Assume that $r_{n_1}^{(a)} = \Omega(r_{n_1}^{c^*})$ as $n \to \infty$ with $0 < c^* < 1/5$ and Conditions 1, 2, 6, 7, and 8 hold, respectively. If we take $n_2 = \Omega(r_{n_1}^{-6c^*})$ and $n_2 = o(r_{n_1}^{-6/5})$ as $n \to \infty$, then we have BAG $\to_d U$ as $n \to \infty$ under H_0 and the asymptotic consistency of the BAGofT under H_1.

For example, suppose that the PTA is a neural network-based method and the number of covariates $p > 46$. Also suppose under H_0, $\pi(\cdot)$ admits a neural network representation, and under H_1, $\pi(\cdot)$ is in the Besov class with the smoothness parameter $\alpha = 2$ (details about the two classes of functions can be found in Yang 1999). According to Yang (1999), typically we have $r_n = O((n/\log n)^{-(p+1)/(4p+2)})$ and $r_n^{(a)} = n^{-2/(4+p)}$. Then, $r_n = O(n^{-1/4})$ and $r_n^{(a)} = \Omega(n^{-1/25})$, so $r_n^{(a)} = \Omega(r_n^{4/25})$. If we set n_2, for example, of the order $n_1^{24(p+1)/(25(4p+2))}$, given the other required conditions for Corollary 1, then the BAGofT asymptotically controls the Type I error probability under H_0 and rejects H_0 with probability going to one under H_1.

4. Practical Guidelines for Implementing the BAGofT

Unlike previous methods in the literature, our approach allows the number of groups to be adaptively chosen, and it may grow when finer partitions are needed to pinpoint the poorly fitted regions. We recommend setting the largest allowed number of groups to $K_{\max} = \sqrt{n_2}$ as a default choice, where $\lfloor a \rfloor$ denotes the largest integer less than or equal to a. We suggest $n_2 = 5\sqrt{n}$ for the training-validation splitting for testing parametric models, which can guarantee enough validation size when $n \geq 100$. In this way, $K_{\max} \to \infty$ as $n \to \infty$, despite that the selected K_n may be small. Our experimental results in Section 5 and the supplementary material show the desirable performance of the default choices under both H_0 and H_1.

4.1. Splitting Ratios and Interpretations in Assessing General Classification Procedures

This subsection includes more details on how to assess the GOF of classification procedures. In practice, the convergence rate r_n under H_0 may not be known. Moreover, when the sample size is finite, the convergence rate r_{n_1} provides limited insight on selecting a suitable splitting ratio. For practical implementations, we advocate considering three splitting ratios where the training set takes 90%, 75%, and 50% of the observations, respectively. The four typical results are given as follows.

Pattern 1: The BAGofT fails to reject H_0 under all the three splitting ratios. The conclusion is that the classification procedure converges quite fast to the underlying conditional probability function, and there is little concern about the lack of fit.

Pattern 2: The BAGofT rejects H_0 only at 50% training. The conclusion is that the classification procedure converges moderately fast, and the procedure fits the data well.

Pattern 3: The BAGofT rejects H_0 at both 50% and 75% and fails to reject at 90%. The conclusion is that the classification procedure converges slowly, but the current sample size is most likely enough for the procedure to fit the data properly.

Pattern 4: The BAGofT rejects H_0 under all the three splitting ratios. The conclusion is that the classification procedure fails to capture the nature of the data generating process.

A caveat is that there may exist "boundary" cases where the 90% training set is still insufficient for the PTA to work well, but the 10% validation set is not enough to identify the weakness of the PTA. In such cases, the failure of rejection may not necessarily be reliable. In general, the BAGofT may have a low power when there is not enough validation data. When the 10% validation set is perceived possibly too small, one possible solution is adding a splitting ratio, for example, 80%, in order to offer more information. Also, note that if $\hat{\pi}(\cdot)$ from the PTA is very sensitive to the sample size and data perturbation, we may fail to observe the gradual change of the rejection results as listed in *Patterns 1–4*. Since unstable procedures are not really reliable anyway, we recommend applying proper stabilization methods to improve the procedure fit first.

4.2. Adaptive Partition for the BAGofT

The asymptotic theory of the BAGofT from the earlier section requires a grouping scheme based on the training set that asymptotically reveals at least one region with $\hat{\pi}_{D_n}(\cdot)$ converging slowly or not converging to $\pi(\cdot)$. In this section, we introduce an adaptive grouping algorithm that may efficiently discover such a region.

The idea of the adaptive grouping is that instead of applying one prescribed partition, we select a partition from a set of partitions based on the training data D_{n_1}. According to Theorems 1 and 3, while protecting the size of the test, we have much flexibility to adaptively select a grouping rule (including the number of groups K_n) as long as it is independent of D_{y_e} conditional on D_{x_e}. Meanwhile, with the adaptive grouping, the power under H_1 is expected to be high.

One way to find a partition to exploit the regions of model misspecification is to fit the deviations (e.g., Pearson residuals) using a nonparametric regression method and choose a partition based on the fitted values. Then, we group the observations with large positive deviations and those with large negative deviations into separate groups to calculate the statistic T for the BAGofT and consequently avoid their cancelation.

In particular, we develop a Random Forest-based adaptive partition scheme as the default choice in our R package 'BAGofT.' It shows excellent performance in our simulation studies. The procedure of the scheme is outlined as follows. On the training set, we first apply the MTA or PTA. We then fit a Random Forest on the training set Pearson residuals and obtain fitted values $\hat{q}_i^{(1)}$, $i = 1, \ldots, n_1$. For different numbers of groups $K = 1, \ldots, K_{\max}$, where $K_{\max} > 2$, we partition $[0, 1]$ into intervals $\{G_1^{(K)}, \ldots G_K^{(K)}\}$ by the K-quantiles of $\{\hat{q}_i^{(1)}\}_{i=1}^{n_1}$, and calculate the statistic

$$\mathcal{B}_K = \sum_{k=1}^{K} \left(\frac{\sum_{\{i:\, \hat{q}_i^{(1)} \in G_k^{(K)}\}} (y_{t,i} - \hat{\pi}_{D_{n_1}}(\boldsymbol{x}_{t,i}))}{\sqrt{\sum_{\{i:\, \hat{q}_i^{(1)} \in G_k^{(K)}\}} \hat{\pi}_{D_{n_1}}(\boldsymbol{x}_{t,i}) \left(1 - \hat{\pi}_{D_{n_1}}(\boldsymbol{x}_{t,i})\right)}} \right)^2 \tag{8}$$

using the training set. We choose the partition $\{G_1^{(K_n)}, \ldots G_{K_n}^{(K_n)}\}$ where K_n is the $K \in 2, \ldots, K_{\max}$ that maximizes $\mathcal{B}_K - \mathcal{B}_{K-1}$. The pseudocode is summarized in Algorithm 1.

Next, we obtain the random forest prediction on the validation set $\hat{q}_i^{(2)}$ $(i = 1, \ldots, n_2)$. We calculate

$$T = \sum_{k=1}^{K_n} \left(\frac{\sum_{\{i:\, \hat{q}_i^{(2)} \in G_k^{(K_n)}\}} R_i}{\sqrt{\sum_{\{i:\, \hat{q}_i^{(2)} \in G_k^{(K_n)}\}} \sigma_i^2}} \right)^2,$$

and then, the p-value statistic BAG in (2).

Note that we may use a set of covariates different from those in the MTA or PTA when applying the random forest learning. For example, we can apply a variable screening to drop some covariates before fitting the classification model or procedure to obtain a parsimonious model or stabilize the fitting algorithm. In this case, our random forest-based adaptive partition may consider all the available covariates to check the GOF. This algorithm also provides some insights on possible misspecifications via the random forest variable importance. Since the random forest is fitted on the Pearson residual of the MTA or PTA, *variables with larger importance are more likely to be associated with the misspecifications*. More details and related

simulations for this algorithm are included in the supplementary material.

Algorithm 1 A default choice of BAGofT adaptive partition

procedure PARTITION($D_{n_1}, K_{max}, parVar$) ▷ *parVar* is the set of variables to construct the partition, which can be different from those in the MTA or PTA (see Section 4.2).
  Fit the MTA or PTA on the set D_{n_1} and calculate the Pearson residual.
  Fit a Random Forest on the Pearson residual with respect to the partition variables *parVar* and obtain the fitted value on the training set $\{\hat{q}_i^{(1)}\}_{i=1}^{n_1}$
  for K in $1, \ldots, K_{\max}$ **do**
    Partition $[0,1]$ by K-quantiles of $\{\hat{q}_i^{(1)}\}_{i=1}^{n_1}$ into $\{G_1^{(K)}, \ldots G_K^{(K)}\}$.
    Calculate $\mathcal{B}_K$ in Equation (8).
  end for
  $K_n \leftarrow \arg\max_{K=2,\ldots,K_{\max}}(\mathcal{B}_K - \mathcal{B}_{K-1})$.
  return $\{G_1^{(K_n)}, \ldots G_{K_n}^{(K_n)}\}$.
end procedure

In high-dimensional settings with many covariates, we have found that a pre-selection used to reduce the number of covariates for the adaptive grouping can help the test performance and save computing cost. We rank the covariates by the correlation distance (Székely, Rizzo, and Bakirov 2007) that measures the dependence relation between the Pearson residual and the covariates, and keep the top ones. More details can be found in the supplementary material.

4.3. Combing Results from Multiple Splittings

Recall that our test is based on splitting the original data into training and validation sets. Due to the randomness of data splitting, we may obtain different test results from the same data. To alleviate this randomness, we can randomly split the data multiple times and appropriately combine the test result from each splitting.

We propose the following procedure. First, we randomly split the data into training and validation sets multiple times and calculate the p-value statistic defined in Equation (2); Second, we calculate the sample mean of the p-value statistic values. Other ways to combine results from multiple splittings include taking the sample median or minimum of the p-value statistic values. It is challenging to derive the theoretical distribution of the statistics from the combined results. Thus, we evaluate the obtained statistic using the bootstrap p-values.

The bootstrap p-value is based on parametric bootstrapping. First, we fit the model using all the data and obtain the fitted probabilities. Second, we generate some bootstrap datasets from the Bernoulli distributions with those fitted conditional probabilities. Third, we calculate the p-value statistic on each of the bootstrap datasets, so these p-value statistics correspond to the case where the MTA or PTA is "correct," Fourth, we compare the p-value statistic from the original data with those from the bootstrap datasets and calculate the bootstrap p-value.

5. Experimental Studies

In the following subsections, we present simulation results to demonstrate the performance of the BAGofT in various settings.

In Section 5.1, we check the performance of the BAGofT in parametric settings and compare it with some existing methods, including the recently proposed Generalized Residual Prediction (GRP) test (Janková et al. 2020). The GRP calculates a test statistic by pivoting the Pearson residuals from the MTA. It has a different focus compared with the BAGofT. First, the GRP test works for generalized linear models (GLM). In contrast, the BAGofT tests general classification models, for example, linear discriminant models and naive Bayes models that do not belong to GLM. Secondly, the GRP test focuses on the cases where the link function of the generalized linear model is correctly specified. The BAGofT can have power against a general deviation of the MTA from the truth. Additionally, when covariates outside the MTA are considered (as mentioned in Section 4.2), the GRP test requires the true model to have the linear effects of these covariates only; the BAGofT can test on other misspecifications, including missing quadratic effects and interactions of the missed covariates. For comparison, we only choose simulation settings that work for both the GRP and BAGofT in this part. A discussion about the required conditions for the BAGofT in the experimental settings is included in the supplementary material.

In Section 5.2, we demonstrate the application of the BAGofT to assess general classification procedures, where we are not aware of any method to compare with.

5.1. Testing on Parametric Models

We choose some commonly studied parametric settings that are similar to those in Pulkstenis and Robinson (2002); Yin and Ma (2013); Canary et al. (2017).

Setting 1. The response is generated from $P(y = 1|x_1, x_2, x_3) = 1/(1 + \exp(-(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)))$, where x_1, x_2, and x_3 are independently generated from Uniform$[-3, 3]$, $\mathcal{N}(0, 1)$, and χ_4^2, respectively. We test the correctly specified model (named Model A) and the model that misses x_3 (named Model B).

Setting 2. The response is generated from $P(y = 1|x_1, x_2) = 1/(1 + \exp(-(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2)))$, where x_1 and x_2 are independently generated from Uniform$[-3, 3]$. We test the correctly specified Model A and Model B that misses the interaction term.

Setting 3. The response is generated from $P(y = 1|x_1, x_2, x_3) = 1/(1+\exp(-(\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_1^2)))$, where x_1, x_2, and x_3 are independently generated from Uniform$[-3, 3]$, $\mathcal{N}(0, 1)$, and χ_2^2, respectively. We test the correctly specified Model A and Model B that misses the quadratic term.

For Model A, we check the null distribution of the BAGofT statistic. For Model B, we compare the power of the BAGofT with the Hosmer-Lemeshow test (Hosmer and Lemeshow 1980), le Cessie-van Houwelingen (CH) test (Le Cessie and Van Houwelingen 1991), and GRP test (Janková et al. 2020). These three tests are fitted by packages *ResourceSelection* (Lele, Keim, and Solymos 2019), *rms* (Harrell Jr 2019), and

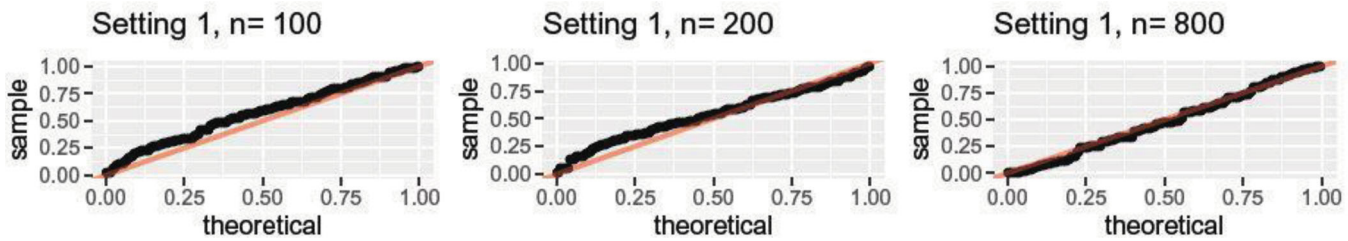

Figure 1. The Q-Q plot of the BAGofT bootstrap *p*-values from Model *A* versus Uniform[0, 1] distribution in Setting 1. The *x*-axis and *y*-axis correspond to the theoretical quantiles and observed sample quantiles, respectively. The red straight line corresponds to the perfect match between the theoretical and observed sample quantiles.

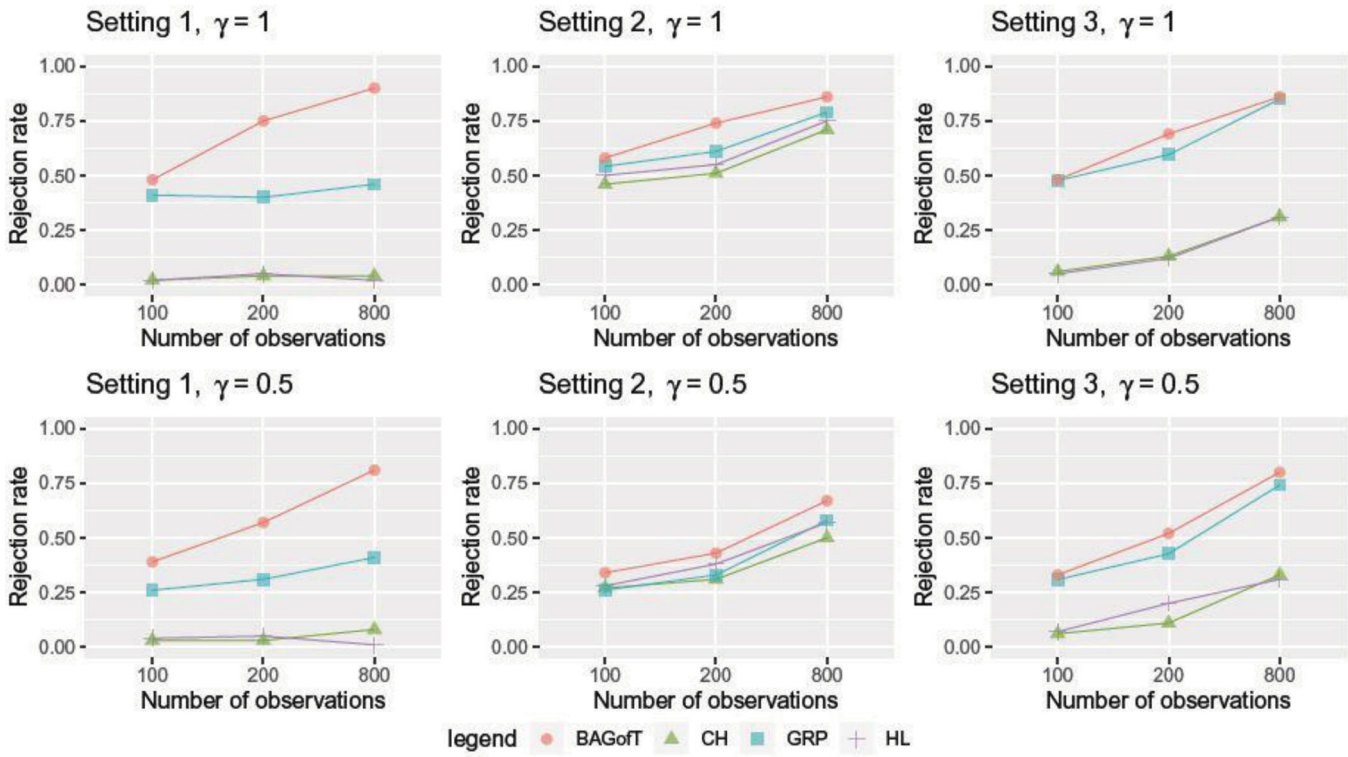

Figure 2. The rejection rates of tests for Model *B* in Settings 1-3. We take standard deviation $\gamma = 1$ or 0.5 for β_3 in Setting 1, Setting 2, and β_4 in Setting 3, respectively. A smaller γ makes it harder to reject. The BAGofT is compared with the HL, CH, and GRP tests. The significance level is 0.05.

GRPtests (Janková et al. 2019), respectively, with their default values. The BAGofT applies 40 data splittings, with all the available covariates considered for the adaptive partition, namely (x_1, x_2, x_3), (x_1, x_2), and (x_1, x_2, x_3) in Settings 1–3, respectively.

To avoid cherry-picking, we independently generate the coefficients from normal distributions with unit standard deviation. Coefficients β_3 in Setting 1 and Setting 2, and β_4 in Setting 3 are generated with mean 1 and others are generated with mean 0. To reflect different degrees of deviation of the MTA from the data generating distribution when testing Model *B*, we consider an additional setting with standard deviation 0.5 for those coefficients generated with mean 1. The other coefficients remain the same as before. The considered sample sizes are 100, 200, and 800, and the testing process in each setting is independently replicated 100 times.

The BAGofT results with the three ways to combine multiple splitting results in Section 4.3 (namely, those based on mean, median, and minimum, respectively) are very close. We thus only present those based on the mean. For Model *A*, the Q-Q plots of the BAGofT *p*-value statistic against Uniform[0, 1] in Setting 1 are shown in Figure 1. We observe that in general, the statistic has a good approximation to Uniform[0, 1] under H_0. When the sample size is small, the simulated Type I error tends to be less than nominal. The results of the other settings are included in the supplementary material, and they show similar results. For Model *B*, the rejection rates of the BAGofT compared with the other tests at the significance level of 0.05 are shown in Figure 2. Due to the random generation of the coefficients, a small portion of the datasets is unbalanced. It caused computation errors for the CH and GRP tests. We dropped these cases when computing the rejection rates. From the results in Figure 2, the BAGofT (in circles) has the best performance in all of the cases. The GRP test (in squares) gets close to the BAGofT in Settings 2 and 3.

We also study the relationship between the number of splittings and the variation of the BAGofT *p*-value statistic. Recall that the purpose of multiple splitting is to obtain a test statistic with smaller variation. The results show that 10 to 20 splittings are usually good enough to get stable results. Additionally, we

check the covariates with the largest variable importance (from the Random Forest fitted on the Pearson residuals) in Settings 1 and 3 when the models are misspecified (Model *B*). Recall that the covariates with large variable importance tend to be the major source of misspecification. Most of the times in our simulation, the missing variable x_3 in Setting 1 has the largest variable importance; x_1 in Setting 2, whose quadratic effect is missing, has the largest variable importance. Additional experimental details on the variations of the statistics and the variable importance are included in the supplementary material.

5.2. Assessing Classification Learning Procedures

In this subsection, we demonstrate the application of the BAGofT to assess classification procedures. We focus on a high-dimensional setting with 1000 covariates and a sample size of 500. A low-dimensional study is included in the supplementary material. The response is generated by the Bernoulli distributions with the following settings.

$$\begin{aligned}
&\textit{Setting 1:} \quad P(y=1|x_1,\ldots,x_{1000}) \\
&\qquad = 1/(1+\exp(-(-6+3\cdot I\{-2<x_1<2\} \\
&\qquad +0.5(x_2+x_3+x_4+x_5)))). \\
&\textit{Setting 2:} \quad P(y=1|x_1,\ldots,x_{1000}) \\
&\qquad = 1/(1+\exp(-(0.5x_1+0.3x_2+0.1x_3+0.1x_4+0.1x_5))).
\end{aligned}$$

The covariates $x_1,\ldots,x_{1000}$ are independently generated from Uniform$[-5,5]$. The PTAs are the logistic regression with LASSO penalty, Random Forest, and XGBoost (Chen and Guestrin 2016).

We first randomly generate the sample data and apply the BAGofT with the three splitting ratios to the PTAs. We apply 20 data splittings, and the adaptive partition is based on all the available covariates $x_1,\ldots,x_{1000}$. The Random Forest is fitted by the package *randomForest* (Liaw and Wiener 2002) with maximum nodes 10. The XGBoost is fitted by the package *xgboost* (Chen et al. 2020) with 25 iterations. The above process is performed with 100 replications, and the results are summarized in Figure 3.

The result of the LASSO logistic regression in Setting 1 belongs to *Pattern 4* since the LASSO logistic fails to capture the nonlinearity in the data-generating model. For Setting 2, it belongs to *Pattern 1* (converging quite fast). The Random Forest has moderate fast or slow convergence speed (*Pattern 2* or *Pattern 3*) in Setting 1. It has a slow convergence speed (*Pattern 3*) or fails to capture the nature of the data generating process (*Pattern 4*) in Setting 2. The Random Forest's overall slow convergence is because its single trees are fitted on some small subsets of the available covariates. As a result, it tends to miss important signals in the sparse setting. The XGBoost converges quite fast (*Pattern 1*) in both settings.

6. Real Data Example

In the following three subsections, we demonstrate the application of the BAGofT by real-world data examples. In Section 6.1, we test a parametric classification model and compare the BAGofT with other methodologies. In Section 6.2, we present a graphical illustration on how the adaptive partition brings an insight on which variables may be responsible for the deficiency of the procedure. In Section 6.3, we apply the BAGofT to assess three classification procedures. We take 20 data splittings and pre-selection size 5 (see Section 4.2) for the BAGofT throughout this section. The significance level is 0.05.

6.1. Tesing Parametric Classification Models: Micro-RNA Data

We consider the study of Shigemizu et al. (2019), where the data is available from the Gene Expression Omnibus (GEO) database with accession number GSE120584. They fitted logistic regressions on micro-RNA data to predict several dementias. Our study focuses on the model that predicts whether a subject has Alzheimer's disease (AD) or not. The data contain $n = 1309$ observations. Shigemizu et al. (2019) selected 78 micro-RNA and computed 10 principal components from the data to fit the

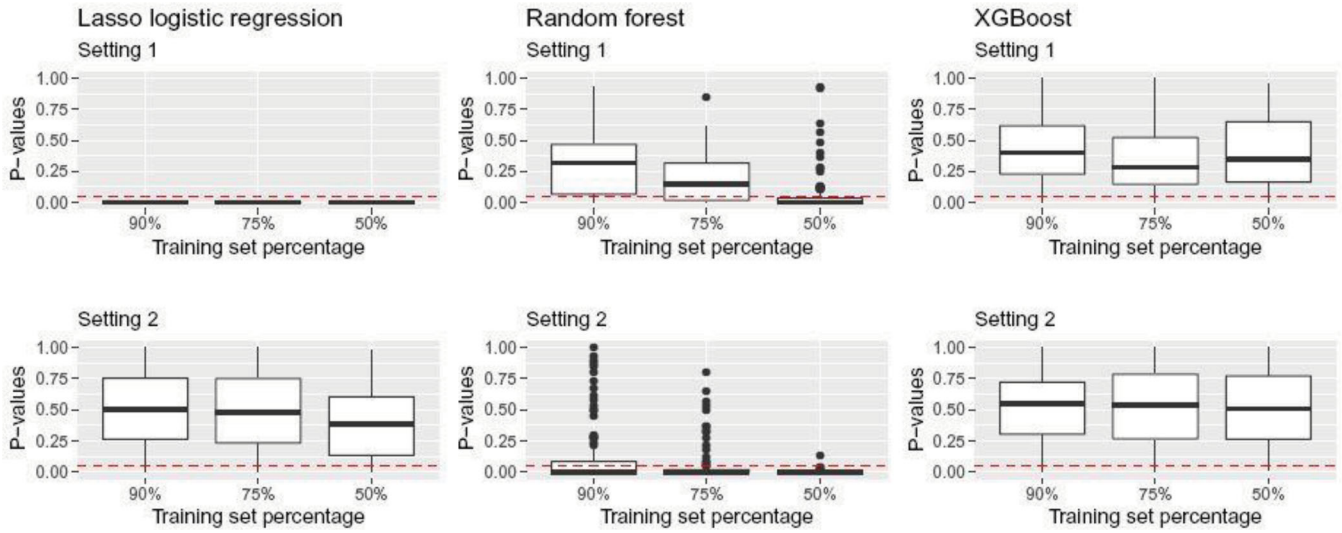

Figure 3. The BAGofT *p*-value boxplots in the high-dimensional settings. The red dashed lines correspond to the 0.05 significance level.

prediction model for AD. We first consider a subset model using the first 7 principal components as the covariates.

$$\textit{Model 1:} \quad \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \mathrm{PC}_1 + \cdots + \beta_7 \mathrm{PC}_7. \quad (9)$$

The available covariates for the BAGofT are the first 20 principal components $\mathrm{PC}_1, \ldots, \mathrm{PC}_{20}$. The bootstrap p-value of the BAGofT is 0. The averaged (random forest) variable importance shows that PC_9 has the largest importance value and is likely to be the major reason for the underfitting.

Next, we add PC_9 to the model and consider:

$$\textit{Model 2:} \quad \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \mathrm{PC}_1 + \cdots + \beta_7 \mathrm{PC}_7 + \beta_9 \mathrm{PC}_9. \quad (10)$$

The p-value from the BAGofT is 0.21. So this model cannot be rejected at the significance level of 0.05.

To compare the performance of the BAGofT with other GOF tests, we also consider the HL, CH, and GRP tests. The results are shown in Table 1. In contrast with the BAGofT, the other tests fail to reject the simpler model, reflecting their lack of power in this case.

6.2. Testing Classification Procedures: Fashion MNIST Data

We consider the Fashion MNIST data (Xiao, Rasul, and Vollgraf 2017), which contain images of different clothes with a pixel size of 28 × 28. We take the first 500 images of trousers and the first 500 images of dresses with a total sample size of 1000. An example snapshot of these images is shown in Figure 4. The PTA is a feed-forward neural network with one hidden layer and one neuron.

Table 1. *P*-values for models from Equations (9) and (10).

Test	HL	CH	GRP	BAG
Model 1	0.42	0.26	0.17	0.00
Model 2	0.17	0.23	0.23	0.15

The BAGofT has a bootstrap p-value 0 in each of the three splitting ratios. It indicates that the neural network fails to capture at least one major aspect from the data (*Pattern 4*). To interpret the testing results, we plot the (random forest) variable importance of the 28 × 28 covariates from the BAGofT (with 90% data for training) in Figure 5. As is remarked in Section 4.2, the covariates with high variable importance are likely to be the major reason for the underfitting. It can be interpreted from Figure 5 that the space between the two legs of the trousers is where the PTA underfits. This is indeed the major difference between the two kinds of clothes.

6.3. Testing Classification Procedures: COVID-19 CT Scans

Coronavirus disease 2019 (COVID-19) has had a massive impact on the world. We consider the data in the study from He et al. (2020), which is available at *https://github.com/UCSD-AI4H/COVID-CT*. The training and test sets contain a total of 339 positive cases and 289 negative cases.

Our study considers assessing classification procedures fitted on the 1000 features generated from the pre-trained deep learning model MobileNetV2 (Sandler et al. 2018). The images are resized into 224 × 224 RGB pixels before entering MobileNetV2. The PTAs are two one-layer neural networks and two XGBoost classifiers. The two neural networks consist of 1 and 7 neurons, respectively. The two XGboost classifiers consist of 10 and 500 base learners, respectively. The details of the PTAs are included in the supplementary material.

The p-values are summarized in Table 2. It can be seen that both the neural network with 1 neuron and XGBoost with 10 base learners are too restrictive to capture the nature of the data (*Pattern 4*). Both the neural network with 7 neurons and XGBoost with 500 base learners belong to *Pattern 1*, and thus handle the data quite well. We also calculate the prediction accuracies of the PTAs by taking 0.5 as the threshold and averaging the accuracies over 100 replications under the three splitting ratios (namely 90%, 75%, and 50%). The result shows that the models not rejected by the BAGofT have accuracies uniformly better than those that are rejected. Note that when assessed

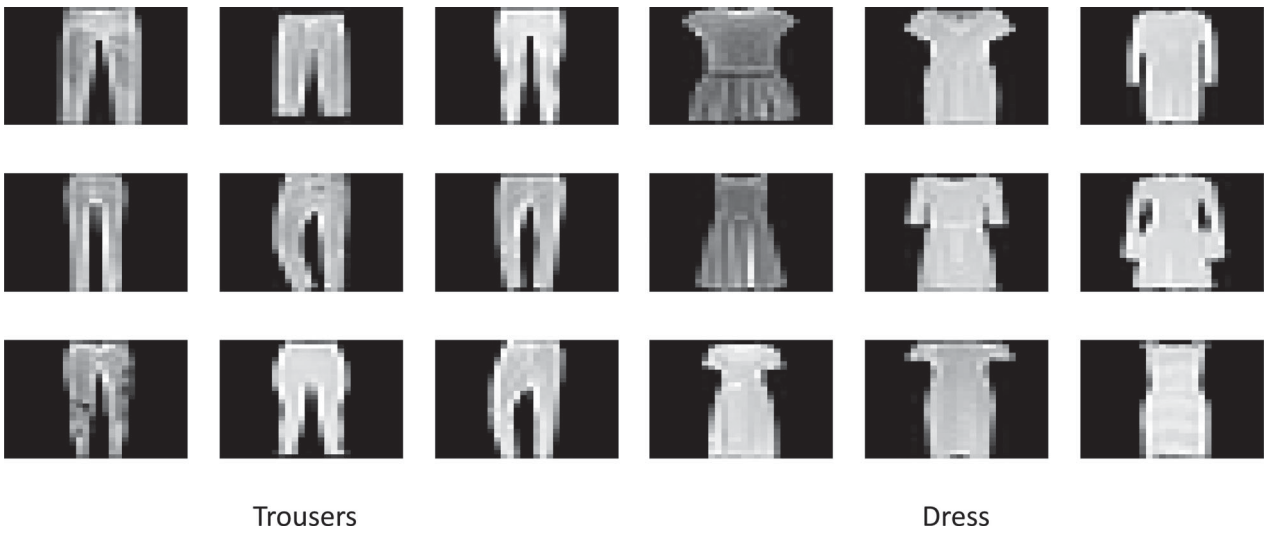

Figure 4. An example of trouser and dress images from the Fashion MNIST data.

Figure 5. Variable importance of the neural network fitted to the Fashion MNIST data. Covariates with higher variable importance are marked by brighter color. The neural network still has room for a major improvement with those highlighted covariates.

Table 2. *P*-values and prediction accuracies from classification procedures fitted on the COVID-19 data (He et al. 2020). NNET-1, NNET-7, XG-10, and XG-500 denote the neural network with 1 neuron, the neural network with 7 neurons, the XGBoost with 10 base learners, and the XGBoost with 500 base learners, respectively.

	P-values			Accuracy		
Splitting ratio	90%	75%	50%	90%	75%	50%
NNET-1	0.03	0.00	0.00	0.71	0.70	0.69
NNET-7	0.62	1.00	0.32	0.72	0.71	0.70
XG-10	0.00	0.00	0.00	0.65	0.65	0.64
XG-500	1.00	0.25	0.60	0.72	0.72	0.70

by prediction accuracies, the neural network with 1 neuron is only slightly worse than the one with 7 neurons. Nevertheless, the BAGofT is able to indicate that the difference in accuracies comes from a systematic defect of the 1-neuron network.

7. Conclusion and Discussion

We have developed a new methodology called the BAGofT to assess the GOF of classification learners. One major novelty is that, unlike the previous methodologies in the literature, it can assess general classification procedures, which is more challenging and has a more extensive application scope than testing parametric models. We have shown both theoretically and experimentally that the BAGofT can effectively reveal different performance patterns of the PTA. Another novelty is the adaptive grouping, which can flexibly expose the MTA or PTA's weaknesses and make the developed tool highly powerful. The adaptive grouping may also be used to interpret which covariates are possibly associated with the underfitting. In the context of assessing parametric models, numerical results have demonstrated the significant advantages of the BAGofT compared with some existing tests, including the popular Hosmer–Lemeshow test.

It is worth emphasizing that the BAGofT has a different usage compared with the assessment tools centered on the classification accuracy. Instead of directly measuring the prediction performance of an MTA/PTA, the BAGofT checks whether it has a detectable systematic issue that leads to slow or non-convergence for the observed data. In one application, the BAGofT can be used by scientists to justify the postulated parametric models and consequently interpret the results on the data-generating mechanism. In another application, data analysts may use the BAGofT to check for systemic defects and make critical business decisions on whether to put more effort on improving an existing MTA/PTA. For many medical and financial applications, it may be valuable to pursue even the smallest improvement of existing methods when we know that they are defective. On the other hand, for other applications where the accuracy at a certain level is fully acceptable, there is no need to perform the BAGofT or other GOF test as long as the accuracy of the MTA/PTA is high enough.

One remaining challenge for the BAGofT is the identification of an overfitted MTA/PTA. When an MTA/PTA is substantially overfitted, the adaptive partition may fail to discover the deviation using the training set because the Pearson residuals may look clean. Nevertheless, in the case of severe overfitting, the chi-squared statistic calculated on the validation set may be able to capture the enlarged variance, and thus the BAGofT may still reject the MTA/PTA. An interesting future direction is to effectively identify large variances from an overfitted MTA/PTA. Another future direction is to extend the BAGofT to the classification problems with $d > 2$ classes. A possible way is to define the statistic T by $\sum_{k=1}^{K_n} \boldsymbol{R_k}^{\mathrm{T}} \boldsymbol{V_k}^{-1} \boldsymbol{R_k}$ where $\boldsymbol{R_k}$ is the sum of differences between the observed response vectors and estimated probabilities from the kth group, and $\boldsymbol{V_k}$ is the estimated covariance matrix for that group. It can be verified by the multivariate Berry–Esseen theorem that $\text{BAG} = 1 - P(\chi^2_{K_n \cdot (d-1)} \leq T | T, K_n)$ has an asymptotic standard uniform distribution under H_0. Nevertheless, a large d brings in computational challenges for the adaptive partition.

The R package "BAGofT" and codes to reproduce the results in Sections 5 and 6 are available at *https://cran.r-project.org/web/packages/BAGofT/index.html*, and *https://github.com/JZHANG4362/BAGofT*, respectively.

Supplementary Materials

In the supplementary file, we provide proofs of the theorems in the main paper and justify/discuss conditions required for the theoretical results. Additional simulations in various settings and real data detailed results are offered as well on performance of BAGofT.

Acknowledgments

The authors thank two anonymous reviewers and the associate editor for their constructive and very helpful comments.

Funding

This article is based upon work supported by the Army Research Laboratory and the Army Research Office under grant number W911NF-20-1-0222, and the National Science Foundation under grant number ECCS-2038603.

References

Bondell, H. D. (2007), "Testing Goodness-of-Fit in Logistic Case-Control Studies," *Biometrika*, 94, 487–495. [1]

Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010), "Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models," *Biometrics*, 66, 1069–1077. [3]

Canary, J. D., Blizzard, L., Barry, R. P., Hosmer, D. W., and Quinn, S. J. (2017), "A Comparison of the Hosmer–Lemeshow, Pigeon–Heyse, and

Tsiatis Goodness-of-Fit Tests for Binary Logistic Regression Under Two Grouping Methods," *Communications in Statistics-Simulation and Computation*, 46, 1871–1894. [6]

Chen, T. and Guestrin, C. (2016), "Xgboost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. [8]

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., and Li, Y. (2020), *xgboost: Extreme Gradient Boosting*, R package version 1.2.0.1. [8]

Fahrmexr, L. (1990), "Maximum Likelihood Estimation in Misspecified Generalized Linear Models," *Statistics*, 21, 487–502. [4]

Farrington, C. P. (1996), "On Assessing Goodness of Fit of Generalized Linear Models to Sparse Data," *Journal of the Royal Statistical Society*: Series B, 58, 349–360. [1]

Harrell Jr, F. E. (2019), *rms: Regression Modeling Strategies*, R package version 5.1-3.1. [6]

He, X., Yang, X., Zhang, S., Zhao, J., Zhang, Y., Xing, E., and Xie, P. (2020), "Sample-Efficient Deep Learning for COVID-19 Diagnosis Based on CT Scans," *medRxiv*. [9,10]

Hosmer, D. W. and Lemeshow, S. (1980), "Goodness of Fit Tests for the Multiple Logistic Regression Model," *Communications in Statistics-Theory and Methods*, 9, 1043–1069. [2,6]

Janková, J., Shah, R. D., Büehlmann, P., and Samworth, R. J. (2019), *GRPtests: Goodness-of-Fit Tests in High-Dimensional GLMs*, R package version 0.1.0. [7]

Janková, J., Shah, R. D., Bühlmann, P., and Samworth, R. J. (2020), "Goodness-of-Fit Testing in High-Dimensional Generalized Linear Models," *Journal of the Royal Statistical Society*, Series B. [6]

Le Cessie, S. and Van Houwelingen, J. C. (1991), "A Goodness-of-Fit Test for Binary Regression Models, Based on Smoothing Methods," *Biometrics*, 1267–1282. [1,6]

Lei, J. (2020), "Cross-Validation With Confidence," *Journal of the American Statistical Association*, 115, 1978–1997. [3]

Lele, S. R., Keim, J. L., and Solymos, P. (2019), *ResourceSelection: Resource Selection (Probability) Functions for Use-Availability Data*, R package version 0.3-5. [6]

Liaw, A. and Wiener, M. (2002), "Classification and Regression by randomForest," *R News*, 2, 18–22. [8]

Liu, Y., Nelson, P. I., and Yang, S.-S. (2012), "An Omnibus Lack of Fit Test in Logistic Regression With Sparse Data," *Statistical Methods & Applications*, 21, 437–452. [2]

Lu, C. and Yang, Y. (2019), "On Assessing Binary Regression Models Based on Ungrouped Data," *Biometrics*, 75, 5–12. [2]

McCullagh, P. (1985), "On the Asymptotic Distribution of Pearson's Statistic in Linear Exponential-Family Models," *International Statistical Review*, 61–67. [1]

Orme, C. (1988), "The Calculation of the Information Matrix Test for Binary Data Models," *The Manchester School*, 56, 370–376. [2]

Osius, G., and Rojek, D. (1992), "Normal Goodness-of-fit Tests for Multinomial Models With Large Degrees of Freedom," *Journal of the American Statistical Association*, 87, 1145–1152. [1]

Pigeon, J. G. and Heyse, J. F. (1999), "An Improved Goodness of Fit Statistic for Probability Prediction Models," *Biometrical Journal*, 41, 71–82. [2]

Pulkstenis, E., and Robinson, T. J. (2002), "Two Goodness-of-fit Tests for Logistic Regression Models With Continuous Covariates," *Statistics in Medicine*, 21, 79–93. [2,6]

Ribeiro, M., Grolinger, K., and Capretz, M. A. (2015), "Mlaas: Machine Learning as a Service," in *Proc. ICMLA*, IEEE, pp. 896–902. [1]

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018), "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520. [9]

Shigemizu, D. et al. (2019), "Risk Prediction Models for Dementia Constructed by Supervised Principal Component Analysis Using miRNA Expression Data," *Communications Biology*, 2, 77. [8]

Stukel, T. A. (1988), "Generalized Logistic Models," *Journal of the American Statistical Association*, 83, 426–431. [1]

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007), "Measuring and Testing Dependence by Correlation of Distances," *The Annals of Statistics*, 35, 2769–2794. [6]

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 1–25. [2,4]

Xian, X., Wang, X., Ding, J., and Ghanadan, R. (2020), "Assisted Learning: A Framework for Multiple-Organization Learning," *Proc. NeurIPS (spotlight)*. [1]

Xiao, H., Rasul, K., and Vollgraf, R. (2017), "Fashion-mnist: A Novel Image Dataset for Benchmarking Machine Learning Algorithms," arXiv:1708.07747. [9]

Xie, X.-J., Pendergast, J., and Clarke, W. (2008), "Increasing the Power: A Practical Approach to Goodness-of-Fit Test for Logistic Regression Models With Continuous Predictors," *Computational Statistics & Data Analysis*, 52, 2703–2713. [2]

Yang, Y. (1999), "Minimax Nonparametric Classification-Part I: Rates of Convergence," *IEEE Transactions on Information Theory*, 45, 2271–2284. [4]

——— (2006), "Comparing Learning Methods for Classification," *Statistica Sinica*, 635–657. [3]

Yin, G., and Ma, Y. (2013), "Pearson-Type Goodness-of-fit Test With Bootstrap Maximum Likelihood Estimation," *Electronic Journal of Statistics*, 7, 412. [2,6]

Yu, Y. and Feng, Y. (2014), "Modified Cross-Validation for Penalized High-Dimensional Linear Regression Models," *Journal of Computational and Graphical Statistics*, 23, 1009–1027. [3]