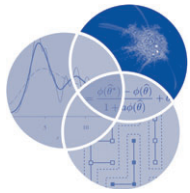
Advanced Review



A survey on Neyman-Pearson classification and suggestions for future research

Xin Tong,[1]* Yang Feng[2] and Anqi Zhao[3]

In statistics and machine learning, classification studies how to automatically learn to make good qualitative predictions (i.e., assign class labels) based on past observations. Examples of classification problems include email spam filtering, fraud detection, market segmentation. Binary classification, in which the potential class label is binary, has arguably the most widely used machine learning applications. Most existing binary classification methods target on the minimization of the overall classification risk and may fail to serve some real-world applications such as cancer diagnosis, where users are more concerned with the risk of misclassifying one specific class than the other. Neyman-Pearson (NP) paradigm was introduced in this context as a novel statistical framework for handling asymmetric type I/II error priorities. It seeks classifiers with a minimal type II error subject to a type I error constraint under some user-specified level. Though NP classification has the potential to be an important subfield in the classification literature, it has not received much attention in the statistics and machine learning communities. This article is a survey on the current status of the NP classification literature. To stimulate readers' research interests, the authors also envision a few possible directions for future research in NP paradigm and its applications.

How to cite this article:
WIREs Comput Stat 2016, 8:64–81. doi: 10.1002/wics.1376

Keywords: Classification, Neyman-Pearson paradigm, plug-in methods, high dimension

INTRODUCTION

Classification has broad applications in various fields, such as biological sciences, medicine, engineering, finance, and social sciences. For example, gene expression data can help predict various types of cancer, analysts can use classification methods to predict whether customers will respond to certain promotions, spam filtering algorithms keep our inbox clean of the junk emails. In general, the aim of classification is to accurately predict discrete outcomes (i.e., class labels) for new observations, on the basis of labeled training data. The development of classification theory, methods and applications has been a dynamic area in statistics and machine learning for more than half a century.[1]

Most existing binary classification methods target on the optimization of the expected classification

*Correspondence to: xint@marshall.usc.edu

[1]Department of Data Sciences and Operations, Marshall School of Business, University of Southern California, Los Angeles, CA, USA

[2]Department of Statistics, Columbia University, New York, NY, USA

[3]Department of Statistics, Harvard University, Cambridge, MA, USA

Conflict of interest: The authors have declared no conflicts of interest for this article.

error R (or 'risk'). The risk is a weighted sum of the type I error R_0 (the conditional probability that the predicted label is 1 given that the true label is 0) and the type II error R_1 (the conditional probability that the predicted label is 0 given that the true label is 1), where the weights are the marginal probabilities of the two class labels. In real-world applications, however, users' priorities for the type I and type II errors may be different from these weights. A representative example of such scenario is the diagnosis of serious disease. Let 1 code the healthy class and 0 code the diseased class. Given that usually

$$\mathbb{P}(Y=1) \gg \mathbb{P}(Y=0),$$

minimizing the overall risk might yield classifiers with small overall risk R (as a result of small R_1) yet large R_0 — a situation quite undesirable in practice given flagging a healthy case incurs only extra cost of additional tests while failing to detect the disease endangers a life. We now demonstrate this point using the neuroblastoma dataset introduced by Ref 2. This dataset contains gene expression profiles of 10707 genes from 246 patients in the German neuroblastoma trials, among which 56 are high-risk patients (labeled 0) and 190 are low-risk patients (labeled 1). The average error rates of PSN[2] (a classifier proposed in Ref 3 under NP paradigm), Gaussian naive Bayes (nb), penalized logistic regression (pen-log), and support vector machines (svm) over 1000 random splits are summarized in Table 1.

Except for PSN[2], all procedures lead to low type II errors and high type I errors. None of the commonly used nb, pen-log, and svm is satisfactory since misdiagnosing a high-risk patient as low-risk (making a type I error) has more severe consequences than the other way around.

One existing approach to asymmetric error control is *cost-sensitive learning*, which allows users to assign two different costs as weights of the type I and type II errors.[4,5] Despite merits of this framework, limitations arise in applications when assigning costs lacks consensus or is morally unacceptable. Also, when users have a specific probabilistic target for the type I/II error control, cost-sensitive learning cannot serve the purpose. Other classification methods targeting small type I errors include the asymmetric support vector machine[6] and the p-value for classification.[7] But like *all previous methods*, they have *no probabilistic guarantee* regarding the type I error bound, resulting in some non-negligible probability of large type I errors. Even if we follow a common practice and tune the empirical type I error (e.g., by adjusting the costs of errors or changing penalty levels) as equal to the targeted level, true type I error of the resulting classifier could actually exceed this level with close to half of the chance!

TABLE 1 | Average Errors for Neuroblastoma Data

Error Type	PSN[2]	Nb	Pen-log	Svm
type I (0 as 1)	.038	.308	.529	.603
type II (1 as 0)	.761	.150	.103	.573

To address such a concern, a novel statistical framework was introduced to control asymmetric errors in binary classification: the Neyman-Pearson (NP) classification paradigm, which seeks a classifier that solves:

$$\min_{\text{type I error} \le \alpha} \text{type II error},$$

where α is a user-specified level, usually a small value (e.g., 5%). The NP paradigm can also prioritize the type II error by symmetry, but we will only discuss the prioritization of type I error in the rest of this paper for presentation consistency.

Although the NP approach has a century-long history in hypothesis testing, it has not been paid much attention in the classification area. This article aims to (1) provide a survey on the current status of NP classification and related literature, and (2) make some suggestions about future research topics in the field. The rest of the article is organized as follows.

- Section *NP Oracle Inequalities* introduces a new theoretical performance measure for NP classification methodology. A large part of our discussion is centered around this measure.
- Section *Empirical Risk Minimization Approach* discusses NP classifiers that follow an empirical risk minimization approach.
- Section *Plug-in Approach* discusses NP classifiers using a plug-in approach with a focus on important theoretical assumptions and challenges in modern high-dimensional settings.
- Section *Selected Topics Related to NP Classification* discusses a few topics that have connections to NP classification.
- Section *Suggestions for Future Research* suggests future research topics in NP classification and NP paradigm in general.

Since we believe that NP classification will undergo a significant development in the next few years, we write the article at this moment to encourage more

researchers to get into the field. Owing to the authors' experience and personal tastes, this survey is inevitably incomplete and biased.

NP ORACLE INEQUALITIES

In this section, we will introduce an important theoretical measure of performance for classifiers in the NP paradigm: the NP oracle inequalities. A few commonly used notations are set up to facilitate our discussion. Let (X, Y) be random variables where $X \in \mathcal{X} \subset \mathbb{R}^d$ is a vector of features and $Y \in \{0, 1\}$ is a class label. A classifier h is a mapping $h : \mathcal{X} \to \{0,1\}$ that returns the predicted class given X. An error occurs when $h(X) \neq Y$, and the binary loss is $\mathbb{I}(h(X) \neq Y)$, where $\mathbb{I}(\cdot)$ denotes the indicator function. The risk is the expected loss with respect to the joint distribution of (X, Y): $R(h) = \mathbb{E}(\mathbb{I}(h(X) \neq Y)) = \mathbb{P}(h(X) \neq Y)$, which can be expressed as a weighted sum of type I and II errors:

$$R(h) = \mathbb{P}(Y = 0)R_0(h) + \mathbb{P}(Y = 1)R_1(h),$$

where $R_0(h) = \mathbb{P}(h(X) \neq Y|Y = 0)$ denotes the type I error, and $R_1(h) = \mathbb{P}(h(X) \neq Y|Y = 1)$ denotes the type II error. While the classical binary classification aims to minimize the risk $R(\cdot)$, the NP classification aims to mimic the NP oracle classifier

$$\phi^* = \arg\min_{\phi: R_0(\phi) \leq \alpha} R_1(\phi),$$

where the user-specified level α reflects a conservative attitude (priority) toward the type I error. Figure 1 shows a toy example that demonstrates the difference between classical and NP classifiers.

Rigollet and Tong[8] argued that, a good classifier $\hat{\phi}$ under the NP paradigm should respect the chosen significance level α. More concretely, two theoretical properties should both be satisfied with high probability.

(I) the type I error constraint is respected, i.e., $R_0(\hat{\phi}) \leq \alpha$.

(II) the excess type II error $R_1(\hat{\phi}) - R_1(\phi^*)$ diminishes with an explicit rate (w.r.t. sample size).

A classifier is said to satisfy NP oracle inequalities if it has properties (I) and (II) simultaneously with high probability. NP oracle inequalities measure theoretical performance of classifiers under the NP paradigm, as well as define a new NP counterpart of the well established oracle inequalities for classifiers in the classical paradigm (see Ref 9 and references within). Recall that, for a classifier $\hat{h}$, the classical oracle inequality insists that with high probability,

$$\text{the excess risk } R\left(\hat{h}\right) - R\left(h^*\right) \text{ diminishes with an explicit rate,}$$

where $h^*(x) = \mathbb{I}(\eta(x) \geq 1/2)$ is the Bayes classifier under the NP paradigm, in which $\eta(x) = \mathbb{E}[Y|X = x] = \mathbb{P}(Y = 1|X = x)$ is the regression function of Y on X.

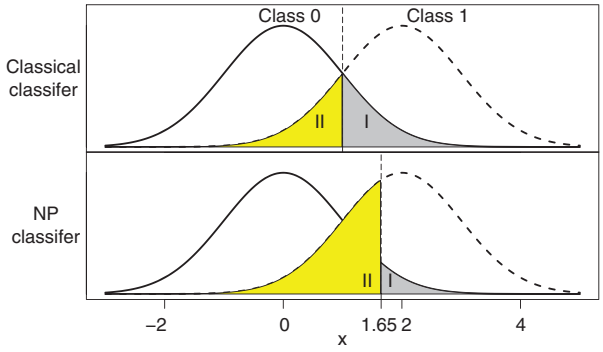


FIGURE 1 | Classical versus NP classifiers in a binary classification example. The true distributions of data x under the two balanced classes are $\mathcal{N}(0,1)$ and $\mathcal{N}(2,1)$ respectively. Suppose that a user prefers a type I error ≤0.05. The classical classifier $\mathbb{I}$ ($x \geq 1$) that minimizes the risk would result in a type I error = 0.16 > 0.05. On the other hand, the NP classifier $\mathbb{I}$ ($x \geq 1.65$) that minimizes the type II error under the type I error constraint (≤0.05) delivers the desirable type I error.

EMPIRICAL RISK MINIMIZATION APPROACH

Existing NP literature can be categorized by following either empirical risk minimization or plug-in approaches. Both approaches are common in the classical classification literature. In this section, we discuss the empirical risk minimization approach to NP classification, and in the next section we investigate the plug-in approach.

NP Paradigm Before 2011

Despite the practical importance of NP classification, a short literature list suffices to summarize the important progress in this field. Cannon et al.[10] initiated the theoretical treatment of the NP classification, and an early empirical study can be found in Ref 11. Several results for traditional statistical learning such as

Probably Approximately Correct (PAC) bounds or oracle type inequalities have been studied in Refs 12,13 in the same framework as the one laid down by Ref 10. Scott[14] proposed performance measures for the NP classification that weigh type I and type II errors in sensible ways. Han et al.[15] transposed several earlier results to the NP classification with a convex loss. All these works follow an empirical risk minimization (ERM) approach, and there is a commonality in this line of literature: a relaxed empirical type I error constraint (i.e., bigger than α) is used in the optimization program, and as a result, the type I error can only be shown to satisfy a relaxed upper bound. In the following, we discuss a few highlights in these papers.

To be consistent with the notations in the literature, we denote in this subsection the NP oracle as $h^* = \arg\min\{R_1(h) : h \in \mathcal{H}, R_0(h) \le \alpha\}$, where $\mathcal{H}$ is some family of classifiers. Cannon et al.[10] established that ERM type classifiers guarantee PAC bounds for fixed tolerance levels $\varepsilon_1, \varepsilon_0 > 0$ as follows. Let $\hat{h}_n$ be a solution to the program

$$\min_{\phi\in\mathcal{H}, \hat{R}_0(\phi)\le\alpha+\varepsilon_0/2} \hat{R}_1(\phi),$$

where $\mathcal{H}$ is a set of classifiers with finite Vapnik-Chervonenkis (VC) dimension V, and where $\hat{R}_0$ and $\hat{R}_1$ denote the empirical type I and type II errors, respectively. It was shown that

1. Under *retrospective sampling* where class 0 sample size n_0 and class 1 sample size n_1 are known before the sample is observed, for any $n(=n_0 + n_1)$,

$$\mathbb{P}\left[\left\{R_0\left(\hat{h}_n\right)-\alpha>\varepsilon_0\right\}\cup\left\{R_1\left(\hat{h}_n\right)-R_1\left(h^*\right)>\varepsilon_1\right\}\right] \le 8n_0^V e^{-n_0\varepsilon_0^2/128} + 8n_1^V e^{-n_1\varepsilon_1^2/128}.$$

2. Under *i.i.d. sampling* in which n_0 and n_1 are unknown until the training sample is observed, if $n \ge \frac{10\sqrt{5}}{\pi_j^2\varepsilon_j^2}$, $\pi_j = \mathbb{P}(Y = j)$, $j = 0, 1$, then

$$\mathbb{P}\left[\left\{R_0\left(\hat{h}_n\right)-\alpha>\varepsilon_0\right\}\cup\left\{R_1\left(\hat{h}_n\right)-R_1\left(h^*\right)>\varepsilon_1\right\}\right] \le 10(2n)^V\left(e^{-\frac{n\pi_0^2\varepsilon_0^2}{640\sqrt{5}}}+e^{-\frac{n\pi_1^2\varepsilon_1^2}{640\sqrt{5}}}\right).$$

Scott and Nowak[13] pointed out that the above bound for *i.i.d. sampling* is substantially larger than that for retrospective sampling, and does not hold for small n. To address this, they proposed an alternative way to derive PAC bounds such that the resulting PAC bounds apply to both sampling schemes for all values of n. This is accomplished by making the tolerance levels ε_0 and ε_1 variable. Specifically, they proved for any $\delta_0, \delta_1 > 0$ and any $n \in \mathbb{N}$:

1. Given a VC class $\mathcal{H}$ with VC dimension V, define

$$\varepsilon_j = \varepsilon_j(n_j,\delta_j,\mathcal{H}) = \sqrt{128\frac{V\log n_j + \log(8/\delta_j)}{n_j}}, \quad j = 0, 1.$$

The classifier $\hat{h}_n$ satisfies

$$\mathbb{P}\left[\left\{R_0\left(\hat{h}_n\right)-\alpha>\varepsilon_0(n_0,\delta_0,\mathcal{H})\right\}\cup\left\{R_1\left(\hat{h}_n\right) - R_1\left(h^*\right)>\in_1(n_1,\delta_1,\mathcal{H})\right\}\right] \le \delta_0+\delta_1.$$

2. Given a finite class $\mathcal{H}$, define

$$\varepsilon_j = \varepsilon_j(n_j,\delta_j,\mathcal{H}) = \sqrt{2\frac{\log|\mathcal{H}| + \log(2/\delta_j)}{n_j}}, \quad j = 0, 1.$$

The classifier $\hat{h}_n$ satisfies

$$\mathbb{P}\left[\left\{R_0\left(\hat{h}_n\right)-\alpha>\varepsilon_0(n_0,\delta_0,\mathcal{H})\right\}\cup\left\{R_1\left(\hat{h}_n\right) - R_1\left(h^*\right)>\varepsilon_1(n_1,\delta_1,\mathcal{H})\right\}\right] \le \delta_0+\delta_1.$$

Scott and Nowak[13] also considered a nested family of classifier classes: $\mathcal{H}^1 \subset \cdots \subset \mathcal{H}^{K(n)}$, and proposed the classifier

$$\widetilde{h}_n = \arg\min_{h\in\mathcal{H}^{K(n)}} \hat{R}_1(h) + \frac{1}{2}\varepsilon_1(n_1,\delta_1,k(h)),$$

$$\text{s.t. } \hat{R}_0(h) \le \alpha + \frac{1}{2}\varepsilon_0(n_0,\delta_0,k(h)),$$

where $k(h)$ is the smallest k such that $h \in \mathcal{H}^k$. The classifier $\widetilde{h}_n$ was shown to satisfy the following theorem.

Theorem 1. *For any n, it holds with probability at least* $1 - (\delta_0 + \delta_1)$,

$$R_0\left(\widetilde{h}_n\right)-\alpha\le\varepsilon_0(n_0,\delta_0,K(n)),$$

$$R_1\left(\widetilde{h}_n\right)-R_1\left(h^*\right)\le\min_{1\le k\le K(n)}\left(\varepsilon_1(n_1,\delta_1,K(n))+\inf_{h\in\mathcal{H}^k,R_0(h)\le\alpha}R_1(h)-R_1\left(h^*\right)\right).$$

Realizing that sometimes, it is necessary to compare classifiers under the NP paradigm, Scott[14] proposed sensible measures that combine type I and type II errors. Denote by f_α^* the classifier that minimizes $R_1(f)$ subject to $R_0(f)\le\alpha$, and set $\beta_\alpha=R_1\left(f_\alpha^*\right)$. Two families of performance measures were proposed, each indexed by a parameter $0<\tau\le\infty$, which reflects the users' trade-off between type I and type II errors:

$$\mathcal{M}^\tau(f)=\tau(R_0(f)-\alpha)_+ +(R_1(f)-\beta_\alpha)_+,$$
$$\mathcal{N}^\tau(f)=\tau(R_0(f)-\alpha)_+ +R_1(f)-\beta_\alpha,$$

where $(x)_+=\max(0,x)$. Both measures penalize any type I error R_0 in excess of α. But compared to $\mathcal{M}^\tau$, $\mathcal{N}^\tau$ encourages small type II error R_1 below the oracle level β_α, potentially at the expense of type I error.

NP Oracle Inequalities Under Convex Loss

We believe that NP oracle inequalities (defined in *NP Oracle Inequalities* section) are an important evaluation metric for classifiers' theoretical performance under the NP paradigm. Bearing these new oracle inequalities as a guideline, Rigollet and Tong[8] proposed a computationally feasible classifier $\widetilde{h}^\kappa$, such that simultaneously with high probability, (*i*) the φ-type I error of $\widetilde{h}^\kappa$, $R_0^\varphi\left(\widetilde{h}^\kappa\right)$, is smaller than α, and (*ii*) the excess φ-type II error of $\widetilde{h}^\kappa$ converges to 0 with explicit rates, where the φ-type I error and φ-type II error are standard convex relaxations of the type I and type II errors by replacing the binary loss with a convex loss φ. Common choices for φ include the hinge loss, the logistic loss, etc.

More concretely, let $\{h_1,\cdots,h_M\}$ be a collection of M base classifiers, and restrict the attention to $\mathcal{H}^{\text{conv}}$, the collection of convex combinations of these base classifiers. The proposed classifier $\widetilde{h}^\kappa$ is a solution to the convex program:

$$\min_{h\in\mathcal{H}^{\text{conv}},\hat{R}_0^\varphi(h)\le\alpha-\kappa/\sqrt{n_0}}\hat{R}_1^\varphi(h),\qquad(1)$$

where $\hat{R}_0^\varphi$ and $\hat{R}_1^\varphi$ are empirical φ-type I/II errors respectively and n_0 and n_1 are sample sizes from class 0 and 1. Note that the empirical φ-type I error is bounded from above by a more stringent level $\alpha-\kappa/\sqrt{n_0}$, which is necessary for controlling the type I error under α. The parameter κ controls how tight we would like to bound the empirical φ-type I error. For example, a large κ means a very stringent bound, which ensures control on the φ -type I error, but will deteriorate the φ-type II error. A careful choice of κ balances φ-type I/II errors as in the following theorem.

Theorem 2. (NP oracle inequalities for $\widetilde{h}^\kappa$ under convex loss; Thm 5 of Tong 2011) *Let φ be Lipschitz on* [−1, 1] *with Lipschitz constant L. Take in (3.1)*

$$\kappa=4\sqrt{2}L\sqrt{\log\left(\frac{2M}{\delta}\right)},$$

where M is the number of base classifiers in $\mathcal{H}^{conv}$. Then under mild regularity conditions, the following hold with probability 1 − 2δ,

$$(I)\ R_0^\varphi\left(\widetilde{h}^\kappa\right)\le\alpha,$$

$$(II)\ R_1^\varphi\left(\widetilde{h}^\kappa\right)-\min_{h\in\mathcal{H}^{\varphi,\alpha}}R_1^\varphi(h)\le C\left(\sqrt{\frac{\log(2M/\delta)}{n_0}}+\sqrt{\frac{\log(2M/\delta)}{n_1}}\right),$$

where $\mathcal{H}^{\varphi,\alpha}=\left\{h\in\mathcal{H}^{\text{conv}},R_0^\varphi(h)\le\alpha\right\}$.

In other words, the classifier $\widetilde{h}^\kappa$ satisfies NP oracle inequalities under a convex surrogate loss. On the technical side, the φ-type I error control is a standard exercise of empirical process theory, while the φ-type II error control involves studying sensitivity of the optimal value to a stochastic constraint set in a convex program.

Type I error R_0 is controlled by α with high probability because $R_0\left(\widetilde{h}^\kappa\right)\le R_0^\varphi\left(\widetilde{h}^\kappa\right)$. On the other hand, the excess φ-type II error bound (inequality (II) of the above theorem) does not imply that the excess type II error of $\widetilde{h}^\kappa$ diminishes to 0. Actually, Rigollet and Tong[8] proved a negative result by

constructing a counter example (Proposition 8 in Ref 8): ERM approaches (using either indicator loss or convex loss function in the optimization program) cannot guarantee diminishing excess type II error as long as one insists the type I error of the proposed classifier be bounded from above by α with high probability. This negative result motivated the study of the NP classification with a plug-in approach in Tong.[16]

PLUG-IN APPROACH

In classical binary classification, plug-in methods that target the Bayes classifier $\mathbb{I}(\eta(x) \geq 1/2)$ have been studied. The earliest works cast doubt on the efficacy of the plug-in approach to classification. For example, Yang[17] showed plug-in estimators cannot achieve excess risk with rates faster than $O(1/\sqrt{n})$ under certain assumptions, while direct methods can achieve fast rates up to $O(1/n)$ under *margin assumption*.[18–21] However, some light was shed on plug-in methods by more recent works starting from Audibert and Tsybakov,[22] which combined a smoothness condition on η with the margin assumption, and showed that plug-in classifiers $\mathbb{I}(\hat{\eta}_n \geq 1/2)$ based on local polynomial estimators can achieve rates faster than $O(1/n)$. The plug-in target under the NP paradigm, however, is not $\mathbb{I}(\eta \geq 1/2)$.

The Oracle Under the NP Paradigm

The oracle classifier under the NP paradigm arises from its close connection to the NP Lemma in statistical hypothesis testing. Hypothesis testing bears strong resemblance to binary classification if we assume the following model. Let P_1 and P_0 be two *known* probability distributions on $\mathcal{X} \subset \mathbb{R}^d$. Let $\zeta \in (0, 1)$ and assume that $Y \sim Bernouli(\zeta)$. Assume further that the conditional distribution of X given Y is denoted by P_Y. Given such a model, the goal of statistical hypothesis testing is to determine whether X was generated from P_1 or from P_0. To this end, we construct a randomized test $\phi : \mathcal{X} \to [0,1]$ and the conclusion of the test based on ϕ is that X is generated from P_1 with probability $\phi(X)$ and from P_0 with probability $1 - \phi(X)$. Two kinds of errors arise: type I error occurs when P_0 is rejected given $X \sim P_0$, and type II error occurs when P_0 is not rejected given $X \sim P_1$. The NP paradigm in hypothesis testing amounts to choosing ϕ that

$$\max. \mathbb{E}[\phi(X)|Y=1], \text{ s.t. } \mathbb{E}[\phi(X)|Y=0] \leq \alpha,$$

where $\alpha \in (0, 1)$ is the significance level of the test. A solution to this constrained optimization problem is called a most powerful test of level α. The NP Lemma gives mild sufficient conditions for the existence of such a test.

Theorem 3. (Neyman-Pearson Lemma) *Let P_0 and P_1 be probability distributions possessing densities q and p respectively with respect to some measure μ. Let $r(x) = p(x)/q(x)$ and C_α be such that $P_0(r(X) > C_\alpha) \leq \alpha$ and $P_0(r(X) \geq C_\alpha) \geq \alpha$. Then for a given level α, the most powerful test of level α is defined by*

$$\phi^*(X) = \begin{cases} 1 & \text{if } r(X) > C_\alpha \\ 0 & \text{if } r(X) < C_\alpha \\ \dfrac{\alpha - P_0(r(X) > C_\alpha)}{P_0(r(X) = C_\alpha)} & \text{if } r(X) = C_\alpha. \end{cases}$$

In other words, under a mild continuity assumption, the plug-in target under the NP paradigm is the **oracle classifier**

$$\phi^*(x) = \mathbb{I}(p(x)/q(x) \geq C_\alpha) = \mathbb{I}(\eta(x) \geq D_\alpha),$$

$$\text{where } D_\alpha = \frac{\mathbb{P}(Y=1)C_\alpha}{\mathbb{P}(Y=1)C_\alpha + \mathbb{P}(Y=0)}.$$

Note that in the classical paradigm, the oracle classifier puts a threshold on the regression function η at precisely 1/2, so plug-in methods do not involve estimating the threshold level. In contrast, NP paradigm faces new challenges, because the threshold level needs to be estimated in addition to the regression function (or the density ratio).

Two Important Theoretical Assumptions

Besides smoothness conditions on density functions, there are two important technical assumptions on the neighborhood of the oracle decision boundary for plug-in classifiers under the NP paradigm.

Definition 1. (margin assumption) *A function $f(\cdot)$ is said to satisfy margin assumption of order $\bar{\gamma}$ with respect to probability distribution P at level C^* if there exists a positive constant M_0, such that for any $\delta \geq 0$,*

$$P\{|f(X) - C^*| \leq \delta\} \leq M_0 \delta^{\bar{\gamma}}.$$

This assumption was first introduced in Polonik.[23] In the classical binary classification framework, Mammen and Tsybakov[18] proposed a similar condition named 'margin condition' by requiring most data to be away from the optimal decision boundary. In the classical classification paradigm, definition 1 reduces to the 'margin condition' by taking $f = \eta$ and $C^* = 1/2$, with $\{x : |f(x) - C^*| = 0\} = \{x : \eta(x) = 1/2\}$ giving the decision boundary of the Bayes classifier. On the other hand, unlike the classical paradigm where the optimal threshold level is known and does not need an estimate, the optimal threshold level C_α in the NP paradigm is unknown and needs to be estimated, suggesting the necessity of having sufficient data around the decision boundary to detect it well. This concern motivated the following condition proposed in Zhao et al.[3] which is an improvement over Tong.[16]

Definition 2. (**detection condition**) *A function $f(\cdot)$ is said to satisfy detection condition of order $\underline{\gamma}$ with respect to P (i.e., $X \sim P$) at level (C^*, δ^*) if there exists a positive constant M_1, such that for any $\delta \in (0, \delta^*)$,*

$$P\{C^* \le f(X) \le C^* + \delta\} \ge M_1 \delta^{\underline{\gamma}}.$$

The detection condition works as an opposite force to the margin assumption, and is basically an assumption on the lower bound of probability. Though a power function was used to keep the lower bound simple and aesthetically similar to the upper bound in margin assumption, any increasing function $u(\cdot)$ on R^+ with $\lim_{x \to 0+} u(x) = 0$ could serve the purpose. Zhao et al.[3] also established the necessity of such a detection condition (in the general sense).

The version of margin assumption and detection condition one should use in the NP paradigm takes $f = r$, $C^* = C_\alpha$, and $P = P_0$ (recall that P_0 is the conditional distribution of X given $Y = 0$). For graphical illustration of these conditions, please refer to Figure 2.

Plug-in Classifiers Under Low-dimensional Settings

Low-dimensional settings refer to the situations where the feature dimensionality d is small and fixed. Under these settings, classical nonparametric estimators can be used to estimate the density ratio p/q or the regression function η. For example, with some proper threshold level estimator $\hat{D}_\alpha$ determined via Vapnik-Chervonenkis theory and the Nadaraya-Watson estimator $\hat{\eta}$, it was shown in Tong[16] that the plug-in classifier $\widetilde{\phi}(x) = \mathbb{I}\left(\hat{\eta}(x) \ge \hat{D}_\alpha\right)$ satisfies NP oracle inequalities.

Theorem 4. (**NP Oracle Inequalities for $\widetilde{\phi}$; Prop 4.2 and Thm 4.1 in Tong 2013**) *Suppose we have access to a mixed i.i.d. sample $\bar{\mathcal{S}} = \{(X_1, Y_1), \ldots, (X_m, Y_m)\}$, and a class 0 sample $\mathcal{S}_0 = \{X_1^-, \ldots, X_n^-\}$. Let $\widetilde{\phi}(x) = \mathbb{I}\left(\hat{\eta}(x) \ge \hat{D}_\alpha\right)$. Assume the regression function η satisfies smoothness conditions, the margin assumption with parameter $\bar{\gamma}$, and the detection condition with parameter $\underline{\gamma}$. In the Nadaraya-Watson estimator*

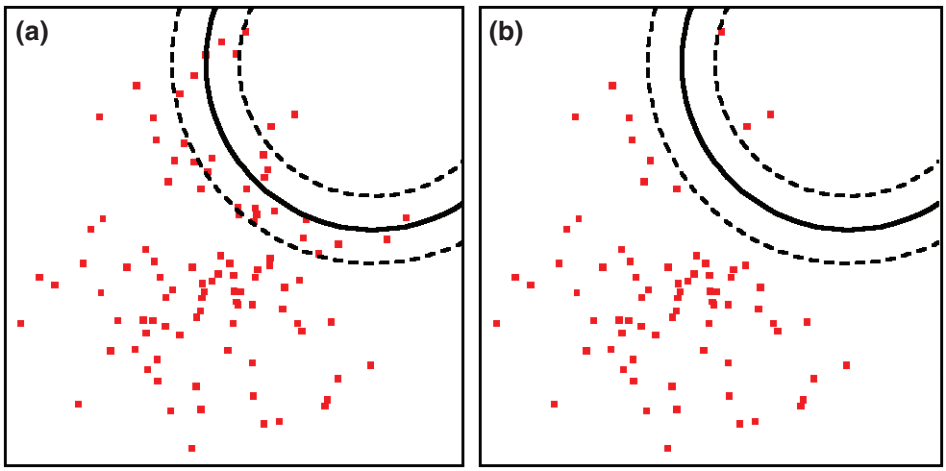


FIGURE 2 | Illustration violation of the margin assumption and detection condition. The solid lines represent the oracle decision boundaries. Subfigure (a) illustrates violation of the margin assumption, and subfigure (b) illustrates violation of the detection condition.

$\hat{\eta}$, where the kernel is β-valid and L'-Lipschitz, take the bandwidth $h = \left(\frac{\log m}{m}\right)^{1/(2\beta+d)}$. *Then there exists a positive constant* $\bar{C}$, *such that for any* $\delta \in (0, 1)$ *and any* $m, n \geq 1/\delta$, *it holds with probability at least* $1 - 3\delta$,

$$(I)\ R_0\left(\widetilde{\phi}\right) \leq \alpha,$$

$$(II)\ R_1\left(\widetilde{\phi}\right) - R_1(\phi^*) \leq \bar{C}\left[\left(\frac{\log n}{n}\right)^{\min\left(\frac{1}{2}, \frac{1+\gamma}{2+\gamma}\right)} + \left(\frac{\log m}{m}\right)^{\frac{\beta(1+\gamma)}{2\beta+d}}\right].$$

A central intermediate step in proving the above result was to derive a high probability bound for $\|\hat{\eta} - \eta\|_\infty$. This uniform deviation bound on Nadaraya-Watson estimators is an interesting result by itself.

Plug-in Classifiers Under High-dimensional Settings

In the big data era, the NP classification framework faces the same 'High Dimension, Low Sample Size' challenge as its classical counterpart does. An overview of general statistical challenges associated with high-dimensionality was given in Hastie et al.[24] and James et al.[25] Despite the NP paradigm's wide big data application potential, Zhao et al.[3] is the first (and so far the only) attempt to construct classifiers satisfying NP oracle inequalities in high-dimensional settings. That paper studied parametric and nonparametric nb models, and proposed a computationally feasible plug-in approach to construct classifiers, which are NOT simple extensions of Tong.[16] The challenge is that one can no longer apply nonparametric estimators blindly in the high-dimensional settings; moreover, a new way to estimate the threshold level is necessary to make the classifiers useful with moderate sample size.

Recall that the NP plug-in target is the oracle $\phi^*(x) = \mathbb{I}(r(x) \geq C_\alpha) = \mathbb{I}(p(x)/q(x) \geq C_\alpha)$ motivated by the NP Lemma. From this oracle, it is clear that next two components should be addressed in proposing any plug-in classifiers in high-dimensional settings under the NP paradigm:

- *build low complexity working models for the density ratio* $r = p/q$. Four model types can be investigated: (I) parametric naive Bayes, (II) nonparametric naive Bayes, (III) linear rules leveraging feature dependence, and (IV) nonlinear rules leveraging feature dependence.

	Independence	Dependence
Linear	(I)	(III)
Nonlinear	(II)	(IV)

- *find a threshold estimate* $\hat{C}_\alpha$ *based on moderate sample size*. Bearing respect to the type I error constraint, Tong[16] required the empirical type I error be bounded from above by $\alpha - t_l$, where $t_l \sim \sqrt{\frac{\log l}{l}}$ and l is the size of class 0 sample not used for estimating $r = p/q$. This approach is of limited practical value except with a large sample size, which is not the case in most modern genetic/genomic applications.

To facilitate the discussion, assume that the available sample contains n i.i.d. observations $\mathcal{S}^1 = \{U_1, \cdots, U_n\}$ from class 1 with density p, and m i.i.d. observations $\mathcal{S}^0 = \{V_1, \cdots, V_m\}$ from class 0 with density q. The samples $\mathcal{S}^1$ and $\mathcal{S}^0$ are decomposed as follows: $\mathcal{S}^1 = \mathcal{S}_1^1 \cup \mathcal{S}_2^1$, and $\mathcal{S}^0 = \mathcal{S}_1^0 \cup \mathcal{S}_2^0 \cup \mathcal{S}_3^0$, where $|\mathcal{S}_1^1| = n_1$, $|\mathcal{S}_2^1| = n_2$, $|\mathcal{S}_1^0| = m_1$, $|\mathcal{S}_2^0| = m_2$, $|\mathcal{S}_3^0| = m_3$. Given this decomposition, Zhao et al.[3] introduced a generic plug-in procedure.

Procedure: NP Plug-in Procedure

Step 1. Use $\mathcal{S}_1^1$, $\mathcal{S}_2^1$, $\mathcal{S}_1^0$, *and* $\mathcal{S}_2^0$ *to construct a density ratio estimate* $\hat{r}$. *The specific use of each subsample will depend on the working models, e.g.,* $\mathcal{S}_1^1$ *and* $\mathcal{S}_1^0$ *are used for independent feature screening when the feature dimensionality is extremely high.*

Step 2. Given $\hat{r}$, *choose a threshold estimate* $\hat{C}_\alpha$ *from the set* $\hat{r}(\mathcal{S}_3^0) = \{\hat{r}(V_{i+m_1+m_2})\}_{i=1}^{m_3}$.

Denote by $\hat{r}_{(k)}(\mathcal{S}_3^0)$ the k th order statistic of $\hat{r}(\mathcal{S}_3^0)$, $k \in \{1, \cdots, m_3\}$. The corresponding plug-in classifier by setting $\hat{C}_\alpha = \hat{r}_{(k)}(\mathcal{S}_3^0)$ is

$$\hat{\phi}_k(x) = \left\{\hat{r}(x) \geq \hat{r}_{(k)}(\mathcal{S}_3^0)\right\}. \tag{2}$$

In rest of this section, we discuss how Zhao et al.[3] modeled and estimated r, and achieved a generic method for choosing k in (2).

Working Models for the Density Ratio

There are many contemporary applications under the broad '$n \ll d$ high-dimensional setting' umbrella. Just as different models are needed for different applications under the classical paradigm, we need an array of working models to handle different applications under the NP paradigm. Zhao et al.[3] discussed the following two models of low complexity. Low complexity is a commonality among many high-dimensional models, because complex models cannot be supported by limited amount of data.

(I) Parametric Naive Bayes

The over simplistic nb models, which ignore all feature dependency, work well in numerous high-dimensional applications. Taking a two class Gaussian model with a common covariance matrix, Bickel and Levina[26] showed that naively carrying out Fisher's discriminant rule performs poorly due to diverging spectra. These authors argued that nb performs better than Fisher's rule in many high-dimensional settings. In addition, even simple models such as nb need to be regularized when we have extremely limited samples. Fan and Fan[27] established the necessity of feature selection for high-dimensional classification problems by showing that even independence rules can be as poor as random guessing due to noise accumulation. When sample size is fairly limited, the (sparse) nb approach is a natural first try for NP classification.

Assuming a two-class Gaussian model $(X|Y=0) \sim \mathcal{N}(\mu^0, \Sigma)$ and $(X|Y=1) \sim \mathcal{N}(\mu^1, \Sigma)$, where $\Sigma = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_d^2)$, Zhao et al.[3] estimated μ^0, μ^1, and Σ using their sample versions $\hat{\mu}^0$, $\hat{\mu}^1$, and $\hat{\Sigma}$. This model is suitable when a linear decision boundary can separate data reasonably well, when correlation among features is low, or when the sample size is so small that one cannot afford to consider more complicated models.

(II) Nonparametric Naive Bayes

The parametric native Bayes model does not allow flexible nonlinear decision boundaries. Hence, Zhao et al.[3] also considered the nonparametric nb model that relaxes the Gaussian assumption and assumes that the conditional distributions of each feature given the class labels are independent:

$$\log \frac{p(x)}{q(x)} = \sum_{j=1}^{d} \log \frac{p_j(x_j)}{q_j(x_j)},$$

where p_j and q_j are the marginal densities of class 1 and 0, respectively, and x_j denotes the j-th component of x, for $j = 1, \cdots, d$. The marginal densities p_j and q_j are approximated by nonparametric estimates $\hat{p}_j$ and $\hat{q}_j$.

(III) and (IV)

Models of types (III) and (IV) will be discussed in the section for future research.

Threshold Estimate $\hat{C}_\alpha$

Leveraging properties of order statistics, Zhao et al.[3] proposed a universal estimate of C_α that works for any given density ratio estimate $\hat{r}$. For a given estimate $\hat{r}$, they found a proper order statistic $\hat{r}_{(k)}(\mathcal{S}_3^0)$ as an estimate of the threshold level C_α, so that type I error of the classifier defined in Eq. (2) is controlled from above by α with high probability.

Proposition 1. For any $\delta_1 \in (0, 1)$ and $k \in \{1, \cdots, m_3\}$, it holds that

$$\mathbb{P}\left(R_0(\hat{\phi}_k) > g(\delta_1, m_3, k)\right) \le \delta_1,$$

where

$$g(\delta_1, m_3, k) = \frac{m_3 + 1 - k}{m_3 + 1} + \sqrt{\frac{k(m_3 + 1 - k)}{\delta_1 (m_3 + 2)(m_3 + 1)^2}}.$$

Let $\mathcal{K} = \mathcal{K}(\alpha, \delta_1, m_3) = \{k \in \{1, \cdots, m_3\} : g(\delta_1, m_3, k) \le \alpha\}$. Proposition 1 implies that $k \in \mathcal{K}(\alpha, \delta_3, m_3)$ is a sufficient condition for the classifier $\hat{\phi}_k$ to satisfy NP Oracle Inequality (I). The next proposition characterizes $\mathcal{K}$, and the smallest $k \in \mathcal{K}$ will ensure small excess type II error for $\hat{\phi}_k$.

Proposition 2. The minimum k that satisfies $g(\delta_1, m_3, k) \le \alpha$ is

$$k_{\min} := \lceil A_{\alpha, \delta_1}(m_3) \cdot (m_3 + 1) \rceil,$$

where $\lceil z \rceil$ denotes the smallest integer larger than or equal to z and

$$A_{\alpha, \delta_1}(m_3) = \frac{1 + 2\delta_1(m_3 + 2)(1 - \alpha) + \sqrt{1 + 4\delta_1(1 - \alpha)\alpha(m_3 + 2)}}{2[\delta_1(m_3 + 2) + 1]}.$$

Zhao et al.[3] proposed the plug-in NP classifier $\hat{\phi}(x) = \hat{\phi}_{(k_{\min})}(x)$, and established the conditions under which it satisfies NP oracle inequalities.

A Critical Intermediate Result

The construction of the threshold estimate $\hat{C}_\alpha$ via order statistics guarantees the type I error bound. To bound the excess type II error, the following intermediate result is critical. This result goes beyond the scopes of Tong[16] and Zhao et al.[3] and can serve as a general strategy for type II error control.

Proposition 3. *Let* $\alpha, \delta_1, \delta_2 \in (0, 1)$. *Assume that the density ratio r satisfies the margin assumption of order* $\bar{\gamma}$ *(with constant* M_0*) and detection condition of order* $\underline{\gamma}$ *(with constant* M_1*), both with respect to distribution* P_0 *at level* C_α. *If* $m_3 \geq \max\{4/(\alpha\delta_1), \delta_1^{-2}, \delta_2^{-2}\}$, *the excess type II error of the classifier* $\hat{\phi}$ *satisfies with probability at least* $1 - \delta_2$,

$$R_1(\hat{\phi}) - R_1(\phi^*) \leq 2M_0 \left[\left(\frac{|R_0(\hat{\phi}) - R_0(\phi^*)|}{M_1} \right)^{1/\underline{\gamma}} + 2\|\hat{r} - r\|_\infty \right]^{1+\bar{\gamma}} + C_\alpha |R_0(\hat{\phi}) - R_0(\phi^*)|.$$

Bounding $|R_0(\hat{\phi}) - R_0(\phi^*)|$, the deviation of type I error of $\hat{\phi}$ away from that of the oracle, is a standard exercise. Hence in view of the above result, to show $\hat{\phi}$ satisfies the NP Oracle Inequalities, the challenging part is to establish a high probability bound for $\|\hat{r} - r\|_\infty$, the uniform deviation of density ratio estimates. In low-dimensional settings, theoretical properties of kernel density estimators have been studied intensively in the literature (see Ref 28 and references therein for a survey). Related results include the convergence in distribution for weighted sup norms derived in Giné et al.[29], and the expected sup-norm loss of multivariate density estimation studied in Lepski[30] using an oracle approach. Tong[16] developed a technical lemma on uniform deviation of kernel density estimates because none of the previous works include such a result, except one with a similar flavor but without explicit constants in the bound.[31] This lemma also contributes to the proof in the nonparametric NP nb settings in Zhao et al.[3]

Numerical Studies

Zhao et al.[3] proposed four classifiers: PN^2 (Parametric NP Naive Bayes), its screening based variant PSN^2, NN^2 (Nonparametric NP nb), and its screening based variant NSN^2. The classifiers PN^2 and PSN^2 are based on model (I), and NN^2, and NSN^2 are based on model (II). The four classifiers were designed for different sample size/dimensionality ratios and different data patterns. The following simulation example and real data example are selected from that paper.

Simulation Example

In the following simulation example, these methods are compared with some commonly used classification methods. Let $(X|Y=1) \sim 0.5\mathcal{N}(a, \Sigma) + 0.5\mathcal{N}(-a, \Sigma)$ and $(X|Y=0) \sim \mathcal{N}(0_d, I_d)$, where $a = (3/\sqrt{10} \cdot 1'_{10}, 0'_{d-10})'$, $\Sigma = \begin{pmatrix} 1/10 \cdot I_{10} & 0 \\ 0 & I_{d-10} \end{pmatrix}$. For $\alpha = 0.05$, the oracle risks are $R_0(\phi^*_\alpha) = 0.05$ and $R_1(\phi^*_\alpha) = 0.027$.

The performance of the screening sub-step of PSN^2 and NSN^2 is shown in Table 2. While both screening methods keep the false positive rates at around 0.05, the parametric screening method (in PSN^2) with *t*-statistic misses almost all signals, while the nonparametric screening method (in NSN^2) using D-statistic[a] does well. This is not surprising since *t*-statistic ranks features by differences in means and the two groups have exactly the same marginal mean and variance across all dimensions.

Figure 3 presents the average errors for different dimensionality *d*. Horizontal axis in each plot of the figure is the logarithm of the sample size *m* (=*n*). The vertical axis of the plots in the first row indicates average type I errors, while that of the plots beneath each of them indicates average type II errors under the same settings. Although all the NP methods work well in controlling the type I error, NSN^2, and NN^2 (based on nonparametric techniques) perform better in terms of the type II error than PSN^2 and PN^2 on non-normal data.

Email Spam Example

This e-mail spam dataset is accessible at https://archive.ics.uci.edu/ml/datasets/Spambase, which contains 4601 observations with 57 features, among which 2788 are class 0 (non-spam) and 1813 are class 1 (spam). Five thousand synthetic features consisting of independent $\mathcal{N}(0,1)$ variables were added to make the problem more challenging. This augmented dataset is split into a training set with 1000 observations from each class and a testing set with the remaining observations. The nonparametric NSN^2 was applied since the sample size is relatively large. The average type I and type II errors over 1000 random splits are shown in Table 3.

To evaluate the flexibility of NSN^2 in terms of prioritized error control, Table 3 also reports the

TABLE 2 | Average Screening Sub-step Performance Summarized over 1000 Independent Simulations at Sample Sizes $m = n = 400$ with Standard Errors in Parentheses

	# of Selected Features		# of Missed Signals		# of False Positive	
d	t-stat	D-stat	t-stat	D-stat	t-stat	D-stat
10	1.76 (1.53)	8.13 (1.83)	8.24 (1.53)	1.87 (1.83)	0 (0)	0 (0)
100	5.93 (3.44)	11.96 (3.57)	9.38 (0.80)	2.34 (1.59)	5.31 (3.17)	4.29 (2.68)
1000	50.69 (9.60)	58.78 (9.87)	9.50 (0.69)	1.26 (1.04)	50.19 (9.51)	50.04 (9.62)

PSN^2 uses t-statistics for screening, and NSN^2 uses D-statistics.

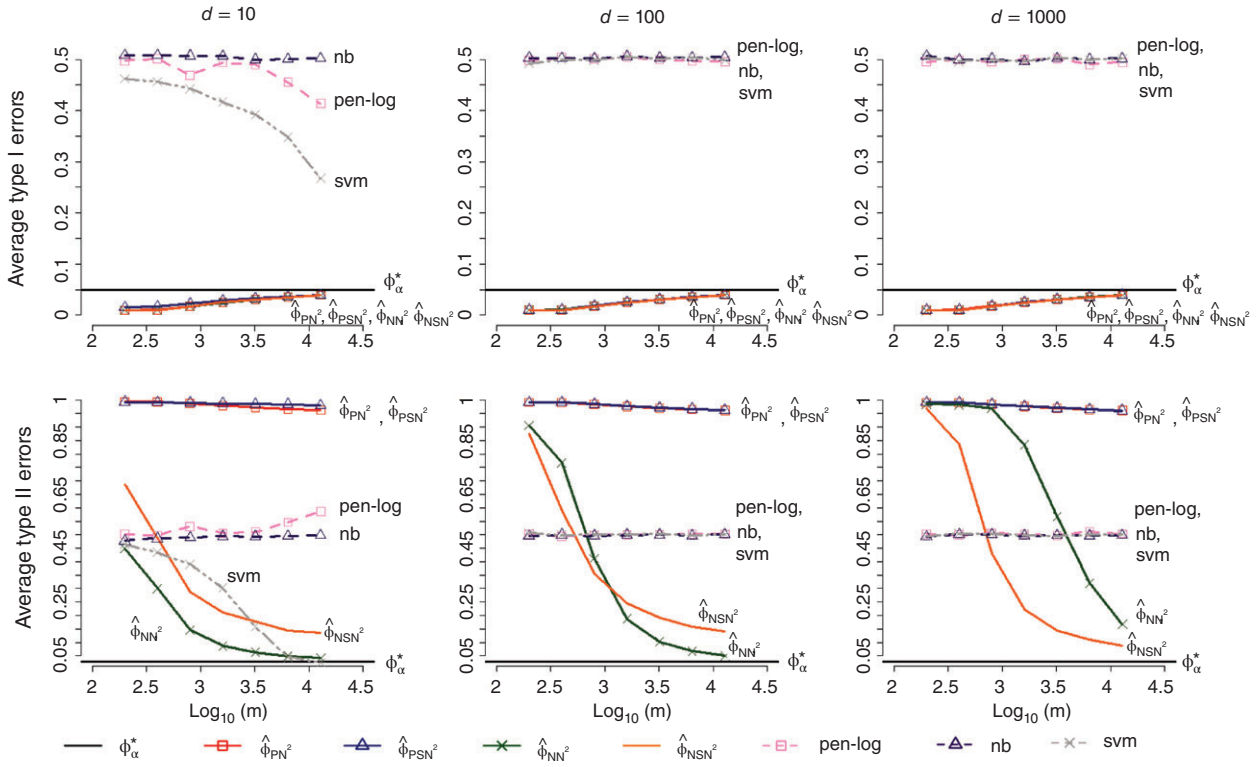


FIGURE 3 | Average error rates of $\hat{\phi}$'s over 1000 independent simulations for each combination of (d, m, n). Error rates are computed as the average of 2000 independent testing data points over 1000 simulations.

performance when the priority is switched to control the type II error below $\alpha = 0.05$. Table 3 demonstrates that NSN^2 can control either type I or type II error depending on the specific need of the practitioner.

SELECTED TOPICS RELATED TO NP CLASSIFICATION

In this section, we connect NP classification to a few other subfields and related papers in machine learning, optimization, and statistics.

Anomaly Detection

NP classification is a useful framework to address anomaly detection problems. In anomaly detection, the goal is to discover outcomes/behaviors that are different from the usual patterns. An unusual behavior is named an *anomaly*. A variety of problems, such as credit card fraud detection, insider trading detection and system malfunctioning diagnosis, fall into this category. There are many approaches to anomaly detection, some serving a specific purpose and others more generic. Modeling techniques include classification, clustering, nearest neighbors, etc. A comprehensive review of anomaly detection literature

TABLE 3 | Average Errors over 1000 Random Splits with Standard Errors in Parentheses

	NSN^2 -R_0	NSN^2 -R_1	Pen-log	Nb	Svm
Type I	.019 (.007)	.488 (.078)	.064 (.007)	.444 (.018)	.203 (.013)
Type II	.439 (.057)	.020 (.009)	.133 (.015)	.054 (.008)	.235 (.017)

The suffix after NSN^2 indicates the type of error it targets to control under α.

is provided by Chandola et al.[32] Earlier review papers include,[33–37] etc.

When we have training data from the normal class, a common approach to anomaly detection is to estimate the normal class density p and try to threshold at a proper level, but this is inappropriate if the anomaly class is far from uniformly distributed. Indeed, to decide whether a certain point is an anomaly, one should consider how likely it is for this point to be normal as opposed to abnormal. The likelihood ratio p/q or the regression function η are good to formalize such a concern. The results in NP classification can be adapted for anomaly detection applications, where the normal sample size is much bigger than the anomaly sample size.

Set Estimation Problems

The plug-in approach to NP classification leads to problems related to density level set estimation (see Ref 38 and references therein), where the task is to estimate $\{x : p(x) > \lambda\}$, for some given level $\lambda > 0$. Density level set estimation has applications in anomaly detection and unsupervised or semi-supervised learning. Plug-in methods for density level set estimation, as opposed to direct methods, do not involve complex optimization procedure, and only amount to thresholding the density estimate at a proper level. The challenges in the NP classification different from those of density level set estimation, such as those in Rigollet and Vert,[38] are two folds. First, the threshold level in the NP setup needs to be estimated, and secondly, NP classification deals with density ratios rather than densities.

NP classification paradigm is also related to the learning of minimum volume set.[39] Given a probability measure P and a reference measure μ, the minimum volume set with mass at least $\beta \in (0, 1)$ is

$$G_\beta^* = \arg\min\{\mu(G) : P(G) \geq \beta, G \text{ measurable}\}.$$

The question of interest is to estimate G_β^* based on independent samples distributed according to the measure P.

The oracle solution to the NP classification with alternative distributions P_0 and P_1 can be re-expressed in terms of acceptance region as

$$G_{1-\alpha}^* = \arg\min\{P_1(G) : P_0(G) \geq 1-\alpha, G \text{measurable}\}.$$

We do not reject (accept) the null if $X \in G_{1-\alpha}^*$. The type I error

$$P_0\left(X \notin G_{1-\alpha}^*\right) = 1 - P_0\left(G_{1-\alpha}^*\right) \leq \alpha$$

satisfies the constraint, whereas the type II error $P_1\left(X \in G_{1-\alpha}^*\right) = P_1\left(G_{1-\alpha}^*\right)$ is minimized. The major difference between minimum volume set estimation problem and NP classification problem is that the reference measure μ in the former is assumed to be known, while P_1 in the latter is unknown.

Semi-supervised Learning

Blanchard et al.[40] developed a general solution to semi-supervised novelty detection by reducing it to the NP classification. Let P_{mixture} denote the marginal distribution of X after integrating out Y. Blanchard et al.[40] made an important observation that the optimal test of size α for hypothesis testing problem

$$H_0 : X \sim P_0, \quad H_{\text{mixture}} : X \sim P_{\text{mixture}},$$

is also optimal, with the same size α, for hypothesis testing problem

$$H_0 : X \sim P_0, \quad H_1 : X \sim P_1.$$

Specifically, for classifier $f : \mathbb{R}^d \to \{0, 1\}$, in addition to the conventional type I/ II error under 0–1 loss $R_y(f) = P_y(f(X) \neq y)$, $y = 0, 1$, define

$$R_{\text{mixture}}(f) = P_{\text{mixture}}\{f(X) = 0\},$$

as the error of misclassifying a sample from the mixture distribution as from P_0. Let

$$R^*_{1,\alpha}(\mathcal{F}) = \inf_{\substack{f \in \mathcal{F} \\ R_0(f) \le \alpha}} R_1(f), \tag{3}$$

$$R^*_{\text{mixture},\alpha}(\mathcal{F}) = \inf_{\substack{f \in \mathcal{F} \\ R_0(f) \le \alpha}} R_{\text{mixture}}(f). \tag{4}$$

Assumption. *For any $\alpha \in (0, 1)$, there exists $f^* \in \mathcal{F}$ such that $R_0(f^*) = \alpha$ and $R_1(f^*) = R^*_{1,\alpha}(\mathcal{F})$,*

Theorem 5. (Thm 1 in Blanchard et al.[40]) *Under Assumption 1, consider any $\alpha \in (0, 1)$, and assume $\pi = \text{IP}(Y = 1) > 0$. Then for any $f \in \mathcal{F}$, the two following statements are equivalent:*

1. $R_{\text{mixture}}(f) = R^*_{\text{mixture},\alpha}(\mathcal{F})$ and $R_0(f) \le \alpha$,
2. $R_1(f) = R^*_{1,\alpha}(\mathcal{F})$ and $R_0(f) = \alpha$.

More generally, let $L_{1,\alpha}(f, \mathcal{F}) = R_1(f) - R^*_{1,\alpha}(\mathcal{F})$ and $L_{\text{mixture},\alpha}(f, \mathcal{F}) = R_{\text{mixture}}(f) - R^*_{\text{mixture},\alpha}(\mathcal{F})$, and assume $\pi > 0$. If $R_0(f) \le \alpha + \varepsilon$ for $\varepsilon > 0$, then

$$L_{1,\alpha}(f, \mathcal{F}) \le \pi^{-1}\{L_{\text{mixture},\alpha}(f, \mathcal{F}) + (1 - \pi)\varepsilon\}.$$

This theorem suggests estimating the solution to Eq. (3) by solving the surrogate problem Eq. (4). It links the NP paradigm not only to semi-supervised novelty detection (SSND) but also to semi-supervised learning problems in general.

Chance Constraint Optimization

It was mentioned in Rigollet and Tong[8] that implementing the NP paradigm for the convexified binary classification bears strong connections with chance constrained optimization. A recent account of such problems can be found in Ben-Tal et al.,[41] Chapter 2 and we refer to this book for references and applications. A chance constrained optimization problem is of the following form:

$$\min_{\lambda \in \Lambda} f(\lambda) \text{ s.t. } \mathbb{P}\{F(\lambda, \xi) \le 0\} \ge 1 - \alpha, \tag{5}$$

where $\xi \in \Xi$ is a random vector, $\Lambda \subset \mathbb{R}^M$ is convex, α is a small positive number and f is a deterministic real valued convex function. Problem (5) can be viewed as a relaxation of robust optimization that solves problems of the form

$$\min_{\lambda \in \Lambda} f(\lambda) \text{ s.t. } \sup_{\xi \in \Xi} F(\lambda, \xi) \le 0,$$

which essentially corresponds to (5) for the case $\alpha = 0$.

In a parallel form of (5), the NP classification paradigm (restrict the search to some convex combination of M base classifiers) can be recast as

$$\min_{\lambda \in \Lambda} R_1(h_\lambda) \text{ s.t. } P_0\{h_\lambda(X) \le 0\} \ge 1 - \alpha, \tag{6}$$

where Λ is the flat simplex of $\mathbb{R}^M$.

Problem (6) differs from (5) in that $R_1(h_\lambda)$ is not a convex function of λ. Replacing $R_1(h_\lambda)$ by the convexified $R^\varphi_1(h_\lambda)$ turns (6) into a standard chance constrained optimization problem:

$$\min_{\lambda \in \Lambda} R^\varphi_1(h_\lambda) \text{ s.t. } P_0\{h_\lambda(X) \le 0\} \ge 1 - \alpha. \tag{7}$$

However, there are two important differences in the NP setting that prevent one from directly using chance constraint optimization techniques, such as Scenario Approach and Bernstein Approximation, to solve (7). First, $R^\varphi_1(h_\lambda)$ is an *unknown* function of λ. Second, NP classification assumes minimum knowledge about P_0, while chance constrained optimization techniques in previous literature need knowledge about the distribution of the random vector ξ.

Classification with Confidence

Classification with confidence is a classification paradigm proposed by Lei.[42] Given $\alpha_0, \alpha_1 \in (0, 1)$, its target optimization problem is formulated as

$$\min_{\substack{C_0 \cup C_1 = \mathcal{X} \\ P_j(C_j) \ge 1 - \alpha_j}} \mathbb{P}(X \in C_0 \cap C_1), \tag{8}$$

and the decision rule is that assigning label 1 in region $C_1 \backslash C_0$, label 0 in region $C_0 \backslash C_1$, and 'ambiguous' in $C_0 \cap C_1$. This paradigm aims to keep the class-specific coverage rates at least $1 - \alpha_j$. The following theorem serves as the oracle for classification with confidence, and plays a role analogous to that of the NP Lemma in the NP paradigm.

Theorem 6. (Thm 1 in Lei[42]) *A solution to the optimization problem (5.8) is*

$$C_0 = \{x : \eta(x) \le t_0\}, \quad C_1 = \{x : \eta(x) \ge t_1\} \cup C_0^c,$$

where $t_0 = t_0(\alpha_0)$ and $t_1 = t_1(\alpha_1)$ are chosen such that $P_0\{\eta(X) \le t_0\} = 1 - \alpha_0$, and $P_1\{\eta(X) \ge t_1\} = 1 - \alpha_1$.

The optimization constraint can also be the overall coverage control, then the problem becomes

$$\min_{C_0 \cup C_1 = \mathcal{X}} \quad \mathbb{P}(X \in C_0 \cap C_1). \\ \pi_0 P_0(C_0) + \pi_1 P_1(C_1) \ge 1 - \alpha$$

SUGGESTIONS FOR FUTURE RESEARCH

On a broader level, NP paradigm is a general formal framework for balancing two conflicting interests. For instance, in addition to the type I/II errors in hypothesis testing and classification, these interests could be in false discovery rate and false negative rate in multiple testing. Moreover, in the same spirit that oracle inequalities in the classical classification paradigm is not the sole criterion in deciding the fate of a classifier, the authors' own proposal of NP oracle inequalities should not be an excuse to exclude useful NP classifiers from designed and practiced. In particular, we think that NP variants of popular classification methods are in demand, but they should not be done in a naive manner such as tuning the empirical type I error to α, which might result in a type I error violation rate close to 50%. To stimulate readers' creative minds, the authors suggest a few possible directions for future research in NP paradigm and its applications.

Leveraging Dependence Under High-dimensional Settings

Zhao et al.[3] considered plug-in classifiers based on nb models under high-dimensional settings. This leaves methods exploiting feature dependency an unexplored territory. Dependence among features is usually an essential characteristic of data,[43] and it can reduce classification error under suitable models and given relative data abundance. The challenge here is that the available sample size is small compared to the feature dimensionality, so the working models for densities should be simple. In particular, the following two working models are worth to consider.

Linear Rules Leveraging Feature Dependence

The sparse versions of Linear Discriminant Analysis (LDA) model

$$(X|Y=0) \sim \mathcal{N}(\mu_0, \Sigma) \text{ and } (X|Y=1) \sim \mathcal{N}(\mu_1, \Sigma),$$

where Σ is a common covariance matrix have been considered in many recent works[44–48] under the classical paradigm. It is worthwhile to consider this model because it is the simplest among all models that take feature dependence into account. Obtaining density ratio estimate $\hat{r}$ under this model involves estimating the precision matrix Σ^{-1}, or the optimal data projection direction $\Sigma^{-1}(\mu_1 - \mu_0)$. Insights can be learned from recent literature on precision/covariance matrix estimation.

Nonparametric Rules Leveraging Feature Dependence

While the nonparametric nb model allows each dimension to enter the decision boundary in a nonlinear fashion, it lacks consideration of feature dependence. The simplest linear structure should be a first try to glue all features together. The following model is a delicate blend of local complexity and global simplicity:

$$\log \frac{p(x)}{q(x)} = \sum_{j=1}^{d} a_j \log \frac{p_j(x_j)}{q_j(x_j)},$$

where p_j and q_j are estimated nonparametrically and the coefficients a_j's are to be learned. This model was proposed by Fan et al.[49] for classical binary classification under high-dimensional settings, and is the backbone of the FANS method in that paper. FANS demonstrates superior data performance under a wide range of spectrums, so we believe it should be interesting to investigate this model under the NP paradigm for applications where feature dependence is significant and linear decision boundaries do not separate data well.

Extension to Multi-class NP Classification

Originating from binary trade-offs, NP classification methods can also be applied to multi-class ($Y \in \{1, \cdots, K\}$, $K \ge 3$) problems in the following two strategies.

- (Strategy 1) Missing class 1 has more severe consequences than missing others. A two step procedure can be executed: first apply NP methods to classify class 1 versus the rest. Stop if a new observation is assigned to class 1. Otherwise, continue and apply a (multi-class) classification algorithm to choose among classes $\{2, \cdots, K\}$.

- (Strategy 2) There is a hierarchical order $(1 > \cdots > K)$ of severity for misclassification. First apply NP methods to 1 versus $\{2, \cdots, K\}$. Stop if a new observation is assigned to class 1. Otherwise, apply NP methods to 2 versus $\{3, \cdots, K\}$. Continue along this line until the observation is assigned a label.

Clearly, NP oracle inequalities are not immediately applicable for multi-class NP classification. New variants should be developed to measure the theoretical performance of multi-class NP classifiers.

Automatic Disease Diagnosis

Automatic disease diagnosis from patients' genomic data is a long-time challenge in cutting-edge clinical research. This task involves a classification problem, where diseases correspond to different classes, and the goal is to predict the diseases that are most likely associated with a patient's genomic sample. Thanks to the rapid development of high-throughput genomic technologies (e.g., microarray and next-generation sequencing), there exists a large amount of disease related genomic data, which can be used as training data in this classification problem. Taking gene expression data as an example, the NCBI (National Center for Biotechnology Information) Gene Expression Omnibus (GEO) currently contains more than 690,000 human gene expression samples, which are related to hundreds of diseases such as heart diseases, mental illnesses, infectious diseases, and various types of cancers.

A novel strategy to tackle automatic disease diagnosis is using NP classification and network-assisted correction. The procedure involves two steps (Figure 4). Step 1: (a) based on public microarray gene expression datasets with 110 Unified Medical Language System (UMLS) class labels (i.e., standardized disease concepts), use NP classification to build a binary classifier for each disease class, and (b) classify a patient's microarray gene expression sample into these disease classes. Step 2: (c) correct the predicted diseases based on the disease taxonomy (network). In Step 1, since the disease classes are nonexclusive (one dataset may have multiple disease class labels), this multi-label classification problem is inherently composed of multiple binary classification problems, where every disease class needs a binary decision. In previous works,[50,51] binary classifiers such as svm and nb classifiers were used for this task, and all disease classes were treated in an interchangeable manner. This has raised an issue: some diseases are more life-threatening than others, as in the example of lung cancer versus athritis. Therefore, it is important to allow doctors to have different levels of conservativeness, i.e., pose different threshold values α on the type I error (missing a disease when a patient in fact has it), for different diseases. Although previous researchers have attempted to address this trade-off between false positives and false negatives in disease diagnosis,[52] their approach lacks a

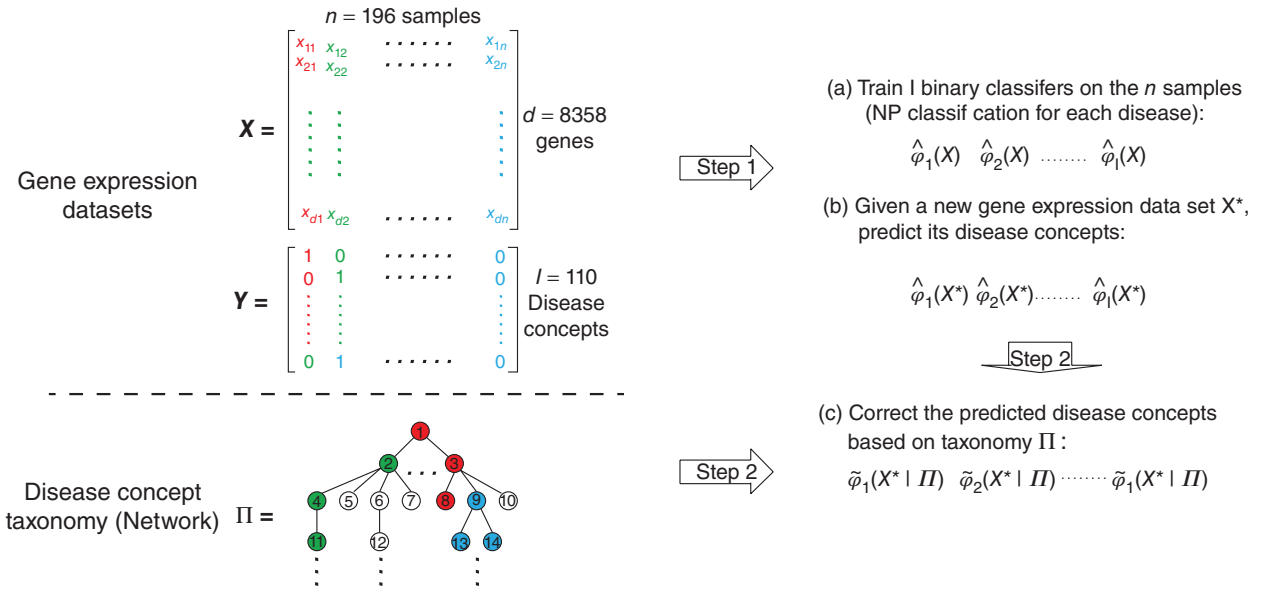


FIGURE 4 | Automatic disease diagnosis via NP classification and network-assisted correction.

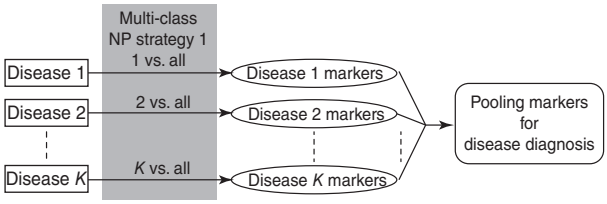


FIGURE 5 | Marker detection via multi-class NP strategy 1.

theoretical guarantee. The NP classifiers can work precisely as a cure for this issue because of their flexibility and theoretical guarantee on the type I error control.

Disease Marker Detection

Multi-class NP classifiers can be used to detect/screen key markers (i.e., genes and genomic features) for disease diagnosis, so as to help researchers gain insights into the molecular pathways and mechanisms of diseases. In previous multiple cancer diagnosis studies,[52–54] to determine which genes should be included as features (markers), classification error of each disease class versus others was used as a criterion. In other words, 'the smallest set' of genes that can result in low classification error for a disease class is retained as key markers of that disease. However, this criterion lacks consideration of the priority of the types I/II errors, and the selected markers for a disease could lead to high false negative rates in the diagnosis of that disease—a dangerous situation for severe diseases such as cancers. Therefore, in the diagnosis of severe diseases, a more reasonable criterion would be the type II error given a pre-specified type I error control. Using the NP classification, key markers would be selected as those leading to a low type II error while retaining the type I error below a threshold (see Figure 5). Markers selected by this new detection scheme will be pooled to make disease prediction and thus increase the sensitivity of disease diagnosis.

Multi-agent Binary Decision (Voting)

Voting is an important social learning theme in democratic societies. When there are two candidates (coded by 0 and 1), a voting problem is similar to a binary classification problem. The difference is that classification can be considered as a single-agent binary decision problem based on n data points, while voting is a multi-agent decision problem in which data exchanges can be considered. This is analogous to the difference between typical statistical inference settings and the multi-agent inference paradigm.[55]

One concrete problem of interest is to derive a finite sample NP version for the renowned *Condorset's Jury Theorem* in Political Science and Political Economics. Condorset's Jury Theorem (CJT) (see Ref 56 and references within) applies to the situation that n voters decide between two alternatives 0 and 1. One of the decisions is 'correct', but they do not know which. Assume that each voter acts independently, and makes the correct decision with probability bigger than 1/2. CJT says, as $n \to \infty$, the probability of the group coming to a correct decision by *majority vote* tends to 1. Variants of CJT have considered dependent voters and/or included finite sample (i.e., fixed n) results.

One can consider a situation where people decide between 0 (the status quo) and 1 (a new alternative), and the choices are not symmetric as the cost of replacing 0 can be high. In this situation, it is reasonable to keep a low probability of mistakenly switching to 1, and the NP paradigm naturally fits in. It would be interesting to develop *new group decision rules* that minimize the probability of wrongly sticking with 0 while keeping the probability of mistakenly switching to 1 at some low level. One can also take into account information exchange among voters before voting, and study the role of network structures in information exchange. Large deviation techniques in statistical learning theory are expected to be useful.

NOTE

[a] The marginal D-statistic for the j-th feature is defined by $D_j = \| \hat{F}_j^0 - \hat{F}_j^1 \|_\infty, \quad j = 1, 2, \cdots, d$

ACKNOWLEDGMENTS

The authors thank the editor, the AE and the referees for their constructive comments that have greatly improve the scope of the article. The research is partially supported by NSF grant DMS-1308566.

REFERENCES

1. Kotsiantis SB, Zaharakis ID, Pintelas PE. Supervised machine learning: a review of classification techniques. *Informatica* 2007, 31:249–268.
2. Oberthuer A, Berthold F, Warnat P, Hero B, Kahlert Y, Spitz R, Ernestus K, König R, Haas S, Eils R, et al. Customized oligonucleotide microarray gene expression "based classification of neuroblastoma patients outperforms current clinical risk stratification. *J Clin Oncol* 2006, 24:5070–5078.
3. Zhao A, Feng Y, Wang L, Tong X. Neyman-Pearson classification under high dimensional settings; 2015. Available at: http://arxiv.org/abs/1508.03106.
4. Elkan C. The foundations of cost-sensitive learning. In: *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'01, San Francisco, CA, USA, 973–978, Morgan Kaufmann Publishers Inc, 2001.
5. Zadrozny B, Langford J, Abe N. Cost-sensitive learning by cost-proportionate example weighting. In: *Proceedings of the Third IEEE International Conference on Data Mining*, ICDM '03, Washington, DC, USA, 435–442, IEEE Computer Society, 2003.
6. Wu S-H, Lin K-P, Chen C-M, Chen M-S. Asymmetric support vector machines: low false-positive learning under the user tolerance. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, New York, NY, USA, 749–757, ACM, 2008.
7. Dümbgen L, Igl B, Munk A. P-values for classification. *Electron J Stat* 2008, 2:468–493.
8. Rigollet P, Tong X. Neyman-Pearson classification, convexity and stochastic constraints. *J Mach Learn Res* 2011, 12:2825–2849.
9. Koltchinskii V. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole dEté de Probabilités de Saint-Flour XXXVIII-2008*, vol. 38. Berlin Heidelberg: Springer-Verlag; 2011.
10. Cannon A, Howse J, Hush D, Scovel C. Learning with the Neyman-Pearson and min-max criteria. *Technical Report LA-UR-02-2951*; 2002.
11. Casasent D, Chen X. Radial basis function neural networks for nonlinear Fisher discrimination and Neyman-Pearson classification. *Neural Netw* 2003, 16:529–535.
12. Scott C. Comparison and design of Neyman-Pearson classifiers. *Manuscript*; 2005.
13. Scott C, Nowak R. A Neyman-Pearson approach to statistical learning. *IEEE Trans Inf Theory* 2005, 51:3806–3819.
14. Scott C. Performance measures for Neyman-Pearson classification. *IEEE Trans Inform Theory* 2007, 53:2852–2863.
15. Han M, Chen D, Sun Z. Analysis to Neyman-Pearson classification with convex loss function. *Anal Theory Appl* 2008, 24:18–28.
16. Tong X. A plug-in approach to Neyman-Pearson classificaiton. *J Mach Learn Res* 2013, 14:3011–3040.
17. Yang Y. Minimax nonparametric classification–part I: rates of convergence. *IEEE Trans Inform Theory* 1999, 45:2271–2284.
18. Mammen E, Tsybakov AB. Smooth discrimination analysis. *Ann Stat* 1999, 27:1808–1829.
19. Tsybakov A. Optimal aggregation of classifiers in statistical learning. *Ann Stat* 2004, 32:135–166.
20. Tsybakov A, van de Geer S. Square root penalty: adaptation to the margin in classification and in edge estimation. *Ann Stat* 2005, 33:1203–1224.
21. Tarigan B, van de Geer S. Classifiers of support vector machine type with ℓ 1 complexity regularization. *Bernoulli* 2006, 12:1045–1076.
22. Audibert J, Tsybakov A. Fast learning rates for plug-in classifiers under the margin condition. *Ann Stat* 2007, 35:608–633.
23. Polonik W. Measuring mass concentrations and estimating density contour clusters-an excess mass approach. *Ann Stat* 1995, 23:855–881.
24. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer-Verlag Inc; 2009.
25. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. New York: Springer-Verlag; 2013.
26. Bickel PJ, Levina E. Some theory for Fisher's linear discriminant function, 'naïve Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* 2004, 10:989–1010.
27. Fan J, Fan Y. High-dimensional classification using features annealed independence rules. *Ann Stat* 2008, 36:2605–2637.
28. Wied D, Weiβbach R. Consistency of the kernel density estimator: a survey. *Stat Pap* 2010, 53:1–21.
29. Giné E, Koltchinskii V, Sakhanenko L. Kernel density estimators: convergence in distribution for weighted sup norms. *Probab Theory Relat Fields* 2004, 130:167–198.
30. Lepski O. Multivariate density estimation under sup-norm loss: oracle approach, adaptation and independence structure. *Ann Stat* 2013, 41:1005–1034.
31. Lei J, Rinaldo A, Wasserman L. A conformal prediction approach to explore functional data. *Ann Math Artif Intell* 2015, 74:29–43.
32. Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey. *ACM Comput Surv* 2009, 09:1–72.

33. Agyemang M, Barker K, Alhajj R. A comprehensive survey of numeric and symbolic outlier mining techniques. *Intell Data Anal* 2006, 6:521–538.
34. Hodge V, Austin J. A survey of outlier detection methodologies. *Artif Intell Rev* 2004, 2:85–126.
35. Markou M, Singh S. Novelty detection: a review-part 1: statistical approaches. *Signal Process* 2003, 12:2481–2497.
36. Markou M, Singh S. Novelty detection: a review-part 2: network-based approaches. *Signal Process* 2003, 12:2499–2521.
37. Patcha A, Park J-M. An overview of anomaly detection techniques: existing solutions and latest technological trends. *Comput Netw* 2007, 12:3448–3470.
38. Rigollet P, Vert R. Optimal rates for plug-in estimators of density level sets. *Bernoulli* 2009, 15:1154–1178.
39. Scott C, Nowak R. Learning minimum volume sets. *J Mach Learn Res* 2006, 7:665–704.
40. Blanchard G, Lee G, Scott C. Semi-supervised novelty detection. *J Mach Learn Res* 2010, 11:2973–3009.
41. Ben-Tal A, Ghaoui L, Nemirovski A. *Robust Optimization*. Princeton, NJ: Princeton University Press; 2009.
42. Lei J. Classification with confidence. *Biometrika* 2014, 2:1–15.
43. Ackermann M, Strimmer K. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics* 2009, 10:1471–2105.
44. Cai T, Liu W. A direct estimation approach to sparse linear discriminant analysis. *J Am Stat Assoc* 2011, 106:1566–1577.
45. Fan J, Feng Y, Tong X. A road to classification in high dimensional space: the regularized optimal affine discriminant. *J R Stat Soc Ser B (Stat Methodol)* 2012, 74:745–771.
46. Mai Q, Zou H, Yuan M. A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika* 2012, 99:29–42.
47. Shao J, Wang Y, Deng X, Wang S. Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann Stat* 2011, 39:1241–1265.
48. Witten D, Tibshirani R. Penalized classification using Fisher's linear discriminant. *J R Stat Soc Ser B (Stat Methodol)* 2012, 73:753–772.
49. Fan J, Feng Y, Jiang J, Tong X. Feature augmentation via nonparametrics and selection (FANS) in high dimensional classification. J Am Stat Assoc. In press.
50. Huang H, Liu C-C, Xianghong Jasmine Z. Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. *Proc Natl Acad Sci USA* 2010, 107:6823–6828.
51. Liu C-C, Hu J, Kalakrishnan M, Huang H, Zhou XJ. Integrative disease classification based on cross-platform microarray data. *BMC Bioinformatics* 2009, 10(Suppl 1):S25.
52. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 2000, 16:906–914.
53. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang C-H, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci USA* 2001, 98:15149–15154.
54. Segal NH, Pavlidis P, Antonescu CR, Maki RG, Noble WS, DeSantis D, Woodruff JM, Lewis JJ, Brennan MF, Houghton AN, et al. Classification and subtype prediction of adult soft tissue sarcoma by functional genomics. *Am J Pathol* 2003, 163:691–700.
55. Fan J, Tong X, Zeng Y. Multi-agent learning in social networks: a finite population learning approach. *J Am Stat Assoc* 2015, 110:149–158.
56. Estlund DM. Opinion leaders, independence, and condorset's jury theorem. *Theor Decis* 1994, 36:131–162.

