*The Annals of Statistics*
2003, Vol. 31, No. 5, 1393–1412


# A CONCRETE STATISTICAL REALIZATION OF KLEINBERG'S STOCHASTIC DISCRIMINATION FOR PATTERN RECOGNITION. PART I. TWO-CLASS CLASSIFICATION

BY DECHANG CHEN, PENG HUANG AND XIUZHEN CHENG

*Uniformed Services University of the Health Sciences, Medical University of South Carolina and George Washington University*

The method of stochastic discrimination (SD) introduced by Kleinberg is a new method in statistical pattern recognition. It works by producing many weak classifiers and then combining them to form a strong classifier. However, the strict mathematical assumptions in Kleinberg [*The Annals of Statistics* **24** (1996) 2319–2349] are rarely met in practice. This paper provides an applicable way to realize the SD algorithm. We recast SD in a probability-space framework and present a concrete statistical realization of SD for two-class pattern recognition. We weaken Kleinberg's theoretically strict assumptions of *uniformity* and *indiscernibility* by introducing *near uniformity* and *weak indiscernibility*. Such weaker notions are easily encountered in practical applications. We present a systematic resampling method to produce weak classifiers and then establish corresponding classification rules of SD. We analyze the performance of SD theoretically and explain why SD is overtraining-resistant and why SD has a high convergence rate. Testing results on real and simulated data sets are also given.

**1. Introduction.** The method of stochastic discrimination (SD) introduced by Kleinberg (1990, 1996) is a new method for solving general problems in statistical pattern recognition. It is fundamentally different from previous methods in the field. The traditionally used techniques in statistical pattern recognition either assume some explicit forms of underlying population density functions and thus require one to estimate parameters, or assume no mathematical structure of the density functions and require one to pursue their estimates nonparametrically. [See, e.g., Duda, Hart and Stork (2001), Fukunaga (1990), McLachlan (1992) and Ripley (1996).] Such a discussion of estimation is not required in SD, yet SD possesses many important properties, such as high convergence rate, high accuracy, overtraining-resistance and ability to handle overlapping classes.

The underlying ideas behind SD were introduced in Kleinberg (1990). Since then, a fair amount of research has been carried out on this method and on variations of its implementation. [See, e.g., Berlind (1994), Chen (1998), Ho (1995, 1998), Ho and Baird (1998), Kleinberg (1996, 2000) and Kleinberg and Ho (1993, 1996).] The results have convincingly shown that stochastic discrimination is a promising area in pattern recognition.

Received April 1998; revised April 2002.
*AMS 2000 subject classifications.* Primary 68T10; secondary 68T05.
*Key words and phrases.* Discriminant function, accuracy, training set, test set.

The approach to establish classification rules of SD is simply described as follows. One first uses resampling techniques to produce a sequence of weak classifiers in light of training data. An individual weak classifier usually performs poorly on the training data, but has high projectability on the test data. Then one averages these weak classifiers to form a strong classifier. This strong classifier not only has good performance on the training data, but also has high projectability on the test data.

The simplest framework of classification rules of SD for two-class problems is described in the following way. Suppose that certain objects to be classified are coming from one of two classes, say class 1 and class 2. A fixed number ($p$) of measurements made on each object form a feature vector $q$. All the $q$'s constitute a finite feature space $F$, a subset of $p$-dimensional Euclidean space $\mathbb{R}^p$. The task is to classify an object after observing its feature vector $q$, which means one needs a classification rule that claims "$q$ comes from class $i$." The goal can be realized by many known methods [see, e.g., Ripley (1996)] with the aid of a training set $\mathbf{TR} = \{TR_1, TR_2\}$, where $TR_i$ is a given random sample of size $n_i$ from class $i$. Completely different from those existing standard methods is the classification rule of SD.

The classification of SD consists of three steps. First, one randomly generates weak classifiers using rectangular regions. In this context, a rectangular region in $\mathbb{R}^p$ is a set of the points $(x_1, x_2, \ldots, x_p)$ such that $a_i \le x_i \le b_i$ for $i = 1, \ldots, p$, where $a_i$ and $b_i$ are real numbers. For simplicity, a rectangular region is denoted by $\prod_{i=1}^{p}(a_i, b_i)$. Let $\Re_1$ be the smallest rectangular region containing $\mathbf{TR}$. For any fixed $\lambda \ge 1$, let $\Re_\lambda$ be the rectangular region in $\mathbb{R}^p$ such that $\Re_\lambda$ and $\Re_1$ have the same center, $\Re_\lambda$ is similar to $\Re_1$ and the "width" of $\Re_\lambda$ along the $x_i$-axis is $\lambda$ times the corresponding width of $\Re_1$. Suppose that $\Re_\lambda = \prod_{i=1}^{p}(L_i, U_i)$. Inside $\Re_\lambda$, a random rectangular region closely related to the training data may be generated as follows: choose a training feature vector $q = (q_1, \ldots, q_p)$ and numbers $l_i$ and $u_i$ such that $L_i \le l_i \le q_i \le u_i \le U_i$ for $i = 1, \ldots, p$. Then form a rectangular region $R = \prod_{i=1}^{p}(l_i, u_i)$. Let $\beta$, $a$ and $b$ be fixed real numbers with $0 < \beta < 1$ and $0 < a \le b \le 1$. An $S$ is a *weak classifier* if $S$ is a union of a finite number of rectangular regions $R$ constructed above such that $a \le r(S, TR_1 \cup TR_2) \le b$ and $|r(S, TR_1) - r(S, TR_2)| \ge \beta$, where for any subsets $T_1$ and $T_2$ of $F$, $r(T_1, T_2)$ denotes the conditional probability $P_F(T_1|T_2) = P_F(T_1 \cap T_2)/P_F(T_2) = |T_1 \cap T_2|/|T_2|$ with $|T|$ representing the cardinality of the set $T$. In this context, $S$ is treated as the intersection $S \cap F$ when $r(S, \cdot)$ is incurred, and $P_F$ represents the uniform probability measure on $F$, under which each element of $F$ has the same probability. Note that $P_F$ has nothing to do with the marginal distribution of the feature vector. The main usage of $P_F$ is to facilitate the counting task related to feature vectors $q$. Figure 1 illustrates a weak classifier for the case when $p = 2$. Using the above process, one generates $t$ independent weak classifiers $S^{(1)}, \ldots, S^{(t)}$, where $t$ is a natural number.

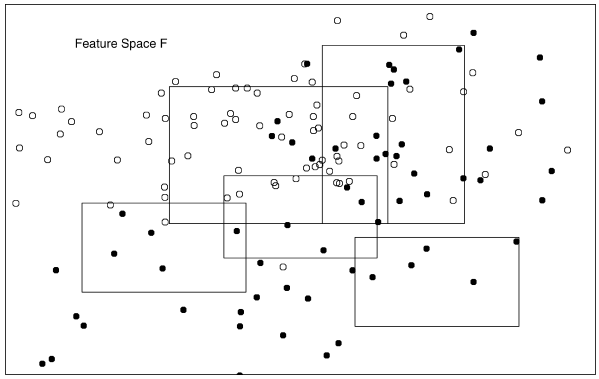


FIG. 1. *An illustration of a weak classifier for two-class classification in the plane. The training data $TR_1$ and $TR_2$ are shown by solid and open figures, respectively. The union of the five rectangular regions, denoted by $S$, is such that $r(S, TR_1) \neq r(S, TR_2)$ and hence is a weak classifier.*

In the second step of SD classification, one combines weak classifiers $S^{(1)}, \ldots, S^{(t)}$ by the central limit theorem. Specifically, for each $q \in F$ one calculates the average $Y(q, \mathbf{S}^t) = (X(q, S^{(1)}) + \cdots + X(q, S^{(t)}))/t$, where $X(q, S)$ is the base random variable, defined to be $(\mathbb{1}_S(q) - r(S, TR_2))/(r(S, TR_1) - r(S, TR_2))$ with $\mathbb{1}_S(q)$ denoting the indicator function. In the last step of the classification, one makes a decision by using the value of $Y(q, \mathbf{S}^t)$. If $Y(q, \mathbf{S}^t) \geq 1/2$, classify $q$ into class 1; otherwise classify $q$ into class 2. The algorithm of SD is outlined in Figure 2.

1. Given $\lambda$, $a$, $b$ and $\beta$, generate $t$ independent weak classifiers.
   (a) Use $\lambda$ to obtain $\Re_\lambda = \prod_{i=1}^{p}(L_i, U_i)$ containing the training data.
   (b) Randomly choose a training feature vector $q = (q_1, \ldots, q_p)$.
   (c) Form $R = \prod_{i=1}^{p}(l_i, u_i)$, where $l_i$ and $u_i$ are randomly selected such that $L_i \leq l_i \leq q_i \leq u_i \leq U_i$.
   (d) Repeat (b) and (c) for a finite number of times to form a union, denoted $S$, of the rectangular regions $R$ such that $r(S, TR_1 \cup TR_2) \in [a, b]$.
   (e) If $|r(S, TR_1) - r(S, TR_2)| \geq \beta$, then retain $S$ as a weak classifier; otherwise discard $S$ and then go to step (d).
   (f) Using the above procedure, obtain $t$ independent weak classifiers $S^{(1)}, S^{(2)}, \ldots, S^{(t)}$.
2. For any $q$ from $F$, evaluate $X(q, S) = \frac{\mathbb{1}_S(q) - r(S, TR_2)}{r(S, TR_1) - r(S, TR_2)}$ at $S^{(1)}, S^{(2)}, \ldots, S^{(t)}$, and then calculate the average
$$Y(q, \mathbf{S}^t) = \frac{X(q, S^{(1)}) + X(q, S^{(2)}) + \cdots + X(q, S^{(t)})}{t}.$$
3. Set a level $t$ classification rule as follows:
   if $Y(q, \mathbf{S}^t) \geq 1/2$, classify $q$ into class 1; otherwise classify $q$ into class 2.

FIG. 2. *SD algorithm.*

This paper mainly carries out a concrete statistical realization of SD for two-class pattern recognition. The strong assumptions of uniformity and indiscernibility in Kleinberg (1996) have been greatly weakened, and the natural notions of near uniformity and weak indiscernibility are introduced. We show why SD classification rules built from training data work well for test data and why SD is overtraining-resistant. We also show why the convergence rate of SD is high. Experimental results on real and simulated data sets are given. Comparisons of SD with other pattern recognition methods are presented.

In Section 2 we study how to establish SD classification rules in detail. Section 3 attacks the issue of the performance of the SD classification rules. Experimental results from simulation and real data sets are given in Section 4. Our conclusion is given in Section 5.

**2. Classification rules.** This section studies the SD algorithm in detail.

2.1. *Weak classifiers.* Condition (e) of the SD algorithm in Figure 2 implies that $r(S, TR_1)$ and $r(S, TR_2)$ are not equal. This fact may be used to define a classification rule for any weak classifier $S$ in the following way. If $r(S, TR_1) > r(S, TR_2)$, assign any $q \in S$ to class 1 and any $q \in S^c$ to class 2. In addition, if $r(S, TR_2) > r(S, TR_1)$, assign any $q \in S$ to class 2 and any $q \in S^c$ to class 1. The error rate on **TR** of such a classification rule can be very high. Thus a weak classifier may be very weak in terms of classification error. Note that the above classification rule of a weak classifier does not take into account the class prior probabilities of the two classes. This will not cause any problem. Combining weak classifiers in SD is done through the central limit theorem and base random variable instead of the direct use of weak classifiers. The base random variable $X$ is used to "separate" the two classes and this separation property is proved without any assumption on the class prior probabilities (see Theorem 1). Additionally, the central limit theorem is used to "amplify" this separation degree (see the first paragraph of Section 2.3).

Denote the collection of all the weak classifiers defined above by **M**, called a *weak classifier space.* Note that **M** depends on **TR**, $\lambda$, $a$, $b$ and $\beta$. For any two given members from **M**, if their intersections with $F$ are the same, then they are equivalent to each other in the sense of classification. This equivalence relationship is implicitly used so that each member of **M** is considered unique.

2.2. *Base random variables.* To connect weak classifiers with feature vectors, we need a special mechanism. This can be done through base random variables $X(\cdot, \cdot)$. As outlined in Figure 2, the base random variable $X$ is the key to forming the classification rule. Theoretically, steps 2 and 3 in the SD algorithm require the central limit theorem to be applied to the sequence $X(q, S^{(1)}), X(q, S^{(2)}), \ldots, X(q, S^{(t)})$, where the feature vector $q$ is fixed. Therefore it is critical to obtain the first and second moments of $X(q, S)$. To calculate

these moments, we apply the usual conditional technique, which is based on a certain partition of $\mathbf{M}$. This partition is made according to the values of $r(S, TR_i)$. The nature of such a partition then leads to the natural near uniformity assumption. Both the partition and assumption enable us to obtain the first and second moments of our base random variables.

Section 2.2.1 gives the strict definition of base random variables, Section 2.2.2 introduces the near uniformity assumption and Section 2.2.3 derives the first and second moments of base random variables.

2.2.1. *Definition of base random variables.* Let $P_{\mathbf{M}}$ denote the uniform probability measure on the weak classifier space $\mathbf{M}$, that is, under $P_{\mathbf{M}}$ each member of $\mathbf{M}$ has the same probability. Define the following function $X$ on $F \times \mathbf{M}$: for $(q, S)$ in $F \times \mathbf{M}$,

$$X(q, S) = \frac{\mathbb{1}_S(q) - r(S, TR_2)}{r(S, TR_1) - r(S, TR_2)}, \tag{1}$$

where $\mathbb{1}_S(q) = 1$ if $q$ is contained in $S$ and 0 otherwise.

Clearly $X$ is a random variable on the probability space $(F \times \mathbf{M}, 2^F \times 2^{\mathbf{M}}, P_F \times P_{\mathbf{M}})$, where $2^F$ and $2^{\mathbf{M}}$ denote the power sets of $F$ and $\mathbf{M}$, respectively. We call $X(\cdot, \cdot)$ a *base random variable.*

The motivation of form (1) can be justified as follows. First note that $\mathbb{1}_S$ is a weak classifier. In fact $\mathbb{1}_S$ functions exactly like $S$ in terms of classification. If $\mathbb{1}_S(q) = 1$, then $q$ is assigned to class 1 if $r(S, TR_1) > r(S, TR_2)$ and assigned to class 2 otherwise. If $\mathbb{1}_S(q) = 0$, then $q$ is assigned to class 2 if $r(S, TR_1) > r(S, TR_2)$ and assigned to class 1 otherwise. For convenience, the standardized version of $\mathbb{1}_S$ is needed. The idea of the standardization is that we seek a transformation of $\mathbb{1}_S(q)$ that has an expectation (restricted to $\mathbf{M}$) close to 1 for $q \in TR_1$ and close to 0 for $q \in TR_2$.

For any fixed $q \in TR_i$, $i = 1, 2$, the expectations $E_{\mathbf{M}}\mathbb{1}_S(q)$ and $E_{\mathbf{M}}r(S, TR_i)$ are identical under the uniformity assumption [Kleinberg (1996), Lemma 3]. Thus informally, $\mathbb{1}_S(q) \approx r(S, TR_1)$ for $q \in TR_1$ and $\mathbb{1}_S(q) \approx r(S, TR_2)$ for $q \in TR_2$. Hence one might try the straightforward (linear) transformation $X(q, S)$ in (1), since informally $X(q, S) \approx 1$ for $q \in TR_1$ and $X(q, S) \approx 0$ for $q \in TR_2$. Additionally, one might guess that such an $X(\cdot, \cdot)$ achieves our goal that $E_{\mathbf{M}}(X(q, S))$ is close to 1 for $q \in TR_1$ and close to 0 for $q \in TR_2$. This is true; see part (a) of Theorem 1 herein.

In summary, the above discussion shows that $X(\cdot, \cdot)$ can be understood as the standardized version of $\mathbb{1}_S(q)$ and $\mathbb{1}_S(q)$ itself is a weak classifier, functioning in the same way as $S$. Note that $X(q, S)$ is symmetric with respect to $TR_1$ and $TR_2$. In fact, by switching $TR_1$ and $TR_2$, we have $X^*(q, S) = [\mathbb{1}_S(q) - r(S, TR_1)]/[r(S, TR_2) - r(S, TR_1)]$ and clearly $X^*(q, S) + X(q, S) = 1$.

2.2.2. *Near uniformity assumption.* In this section, we first give a partition of $\mathbf{M}$ according to the values of $r(S, TR_i)$ and then introduce the near uniformity assumption on the partition. Intuitively, this near uniformity says that each component of the partition of $\mathbf{M}$ is "almost uniformly spread over the training set."

Let $\mathbf{x} = (x_1, x_2)$ be a pair of real numbers with $0 \le x_i \le 1$ for $i = 1, 2$, and define

$$\mathbf{M}_{\mathbf{x}} = \{S : S \in \mathbf{M},\ r(S, TR_i) = x_i \text{ for } i = 1, 2\}.$$

It is easy to see that there exist pairs $\mathbf{x}^{(l)} = (x_1^{(l)}, x_2^{(l)})$ with rational components, $l = 1, 2, \ldots, d$, such that the $\mathbf{x}^{(l)}$'s are different from each other, none of the $\mathbf{M}_{\mathbf{x}^{(l)}}$'s is empty and

$$\mathbf{M} = \bigcup_{l=1}^{d} \mathbf{M}_{\mathbf{x}^{(l)}}. \tag{2}$$

Equation (2) gives a partition of $\mathbf{M}$ into $d$ disjoint sets $\mathbf{M}_{\mathbf{x}^{(1)}}, \mathbf{M}_{\mathbf{x}^{(2)}}, \ldots, \mathbf{M}_{\mathbf{x}^{(d)}}$. Given $l$, all the $S$'s in $\mathbf{M}_{\mathbf{x}^{(l)}}$ cover the same number of feature vectors in $TR_i$.

We now want to impose some natural condition on the partition in (2). Before proceeding with this condition, we motivate it by the following discussion of **TR**. We have $N = n_1 + n_2$ feature vectors in **TR**, where $n_i$ is the size of $TR_i$ $(i = 1, 2)$. Suppose $z_\nu^{(1)} < z_\nu^{(2)} < \cdots < z_\nu^{(k_\nu)}$ is the ordered list of the distinct $\nu$th coordinates (in $\mathbb{R}^p$) for all the feature vectors in **TR**, $\nu = 1, 2, \ldots, p$. For each fixed $\nu$, choose numbers $h_{\nu j}$ such that $z_\nu^{(j)} < h_{\nu j} < z_\nu^{(j+1)}$ for $j = 1, 2, \ldots, k_\nu - 1$. Then the hyperplanes (in $\mathbb{R}^p$) $z_\nu = h_{\nu j}$ divide $\mathbb{R}^p$ into $c = \prod_{\nu=1}^{p} k_\nu$ mutually exclusive and exhaustive subsets $R_1, R_2, \ldots, R_c$ so that each $R$ contains at most one feature vector from **TR**. Consider the following scenario. Let $\mathbf{M}'$ be the set such that $S$ belongs to $\mathbf{M}'$ iff $S$ is a finite union of members in $\{R_1, R_2, \ldots, R_c\}$. Then for the partition (2) associated with $\mathbf{M}'$, we have the conditional probability equality $P_{\mathbf{M}'}(\{S : q \in S\} | \mathbf{M}'_{\mathbf{x}^{(l)}}) = x_i^{(l)}$, for any $l$ $(l = 1, \ldots, d)$, $i$ $(i = 1, 2)$ and $q \in TR_i$. In fact, suppose $x_i^{(l)} = c_{li}/n_i$, where $c_{li}$ is a positive integer. Fix $l$ and consider the action of drawing an $S$ from $\mathbf{M}'_{\mathbf{x}^{(l)}}$ such that the size of $S \cap TR_i$ equals $c_{li}$. It then follows that for any fixed $q \in TR_i$, $\mathbb{1}_S(q)$ is distributed as Bernoulli$(x_i^{(l)})$ and thus the desired conditional probability equality is obtained.

Turning back to our original weak classifier space $\mathbf{M}$, we note that each member of $\mathbf{M}$ is determined by $\lambda$, $a$, $b$ and $\beta$. This restriction usually does not lead to the above conditional probability equality if no correction term is added. Thus the following postulation is natural:

NEAR UNIFORMITY ASSUMPTION. There exists some positive function $\varepsilon(q)$ of $q$, dependent on $\lambda$, $a$, $b$, $\beta$ and **TR**, such that for any $l$ $(l = 1, \ldots, d)$, $i$ $(i = 1, 2)$ and $q \in TR_i$, the conditional probability

$$P_{\mathbf{M}}\big(\{S : q \in S\} | \mathbf{M}_{\mathbf{x}^{(l)}}\big) = x_i^{(l)} + o_{li}(q), \tag{3}$$

where $|o_{li}(q)| \le \varepsilon(q)$.

If all the $o_{li}$ were 0, then (3) would imply that the probability that a feature vector $q_1$ in $TR_i$ is captured by a weak classifier in $\mathbf{M}_{\mathbf{x}^{(l)}}$ is equal to the probability that a feature vector $q_2$ in $TR_i$ is captured by a weak classifier in $\mathbf{M}_{\mathbf{x}^{(l)}}$. Hence, $\mathbf{M}_{\mathbf{x}^{(l)}}$ would be "uniformly spread over training set $TR_i$," which is the idea behind the uniformity assumption in Kleinberg (1996). However, in reality $o_{li}$ may not be 0. Thus, intuitively, (3) tells us that $\mathbf{M}_{\mathbf{x}^{(l)}}$ is just "near uniformly spread over each of $TR_1$ and $TR_2$."

Kleinberg (1996) mentioned the phenomenon of near uniformity, but did no further analysis. Some research results under the assumption of near uniformity can be found in Chen (1998). We will use NUA($\varepsilon$) to refer to the above near uniformity assumption.

2.2.3. *Moments of base random variables.* By utilizing the near uniformity assumption we can establish the following results on the first and second moments of base random variables.

THEOREM 1. *For any given $q$, let $E_{\mathbf{M}}(X(q,S))$ and $\mathrm{Var}_{\mathbf{M}}(X(q,S))$ denote the expectation and variance, respectively, restricted to $\mathbf{M}$, of $X(q,S)$. Assume* NUA($\varepsilon$) *for* $\mathbf{M}$. (a) $E_{\mathbf{M}}(X(q,S)) = 1 + \tau_1(q)$ *if* $q \in TR_1$ *and* $= \tau_2(q)$ *if* $q \in TR_2$; (b) $\mathrm{Var}_{\mathbf{M}}(X(q,S)) \le (1 + 4\varepsilon(q))/(4\beta^2) - \tau_1^2(q) - 2\tau_1(q)$ *for* $q \in TR_1$ *and* $\le (1 + 4\varepsilon(q))/(4\beta^2) - \tau_2^2(q)$ *for* $q \in TR_2$, *where all the* $\tau$*'s satisfy* $|\tau| \le \varepsilon(q)/\beta$.

PROOF. (a) From (2), $\mathbf{M} = \bigcup_{l=1}^{d} \mathbf{M}_{\mathbf{x}^{(l)}}$. Let $p_l = P_{\mathbf{M}}(\mathbf{M}_{\mathbf{x}^{(l)}})$. For any given $q \in TR_i$ $(i = 1, 2)$,

$$
\begin{aligned}
E_{\mathbf{M}}X(q,S) &= \sum_{l=1}^{d} E_{\mathbf{M}}\big(X(q,S)|\mathbf{M}_{\mathbf{x}^{(l)}}\big)p_l \\
&= \sum_{l=1}^{d}\left[\frac{E_{\mathbf{M}}(\mathbb{1}_S(q)|\mathbf{M}_{\mathbf{x}^{(l)}}) - x_2^{(l)}}{x_1^{(l)} - x_2^{(l)}}\right]p_l \qquad (4)\\
&= \sum_{l=1}^{d}\left[\frac{x_i^{(l)} + o_{li}(q) - x_2^{(l)}}{x_1^{(l)} - x_2^{(l)}}\right]p_l,
\end{aligned}
$$

where the last equality comes from NUA($\varepsilon$). Note that $|x_1^{(l)} - x_2^{(l)}| \ge \beta$ for $l = 1, \ldots, d$. Set $u_{li}(q) = o_{li}(q)p_l/(x_1^{(l)} - x_2^{(l)})$. We have $|u_{li}(q)| \le \varepsilon(q)p_l/\beta$. Thus from (4), part (a) of the theorem is established, where $\tau_i(q) = \sum_{l=1}^{d} u_{li}(q)$.

(b) For any given $q \in TR_i$, we have

$$
\begin{aligned}
E_{\mathbf{M}}[X(q,S)]^2 &= \sum_{l=1}^{d} E_{\mathbf{M}}\left(\left(\frac{\mathbb{1}_S(q)-x_2^{(l)}}{x_1^{(l)}-x_2^{(l)}}\right)^2 \Big| \mathbf{M}_{\mathbf{x}^{(l)}}\right) p_l \\
&= \sum_{l=1}^{d} \frac{(1-2x_2^{(l)})P_{\mathbf{M}}(\{S: q\in S\}|\mathbf{M}_{\mathbf{x}^{(l)}}) + (x_2^{(l)})^2}{(x_1^{(l)}-x_2^{(l)})^2} p_l \qquad (5) \\
&= \sum_{l=1}^{d} \frac{x_i^{(l)}(1-2x_2^{(l)}) + (x_2^{(l)})^2 + (1-2x_2^{(l)})o_{li}(q)}{(x_1^{(l)}-x_2^{(l)})^2} p_l.
\end{aligned}
$$

Note that $|1-2x| \le 1$ and $x(1-x) \le 1/4$ for $0 \le x \le 1$. Therefore if $i=1$, (5) gives

$$
\begin{aligned}
E_{\mathbf{M}}[X(q,S)]^2 &= 1 + \sum_{l=1}^{d} \frac{x_1^{(l)}(1-x_1^{(l)}) + (1-2x_2^{(l)})o_{l1}(q)}{(x_1^{(l)}-x_2^{(l)})^2} p_l \\
&\le 1 + \frac{1+4\varepsilon(q)}{4\beta^2},
\end{aligned}
$$

and if $i=2$, a similar argument yields the bound $(1+4\varepsilon(q))/(4\beta^2)$. Part (b) now follows from the above second moments and part (a). □

When the $|\tau_i|$'s in Theorem 1 are small, the expectation of $X(q,S)$ ($q$ fixed) is close to 1 if $q$ is in $TR_1$ and close to 0 if $q$ is in $TR_2$. In a sense, $E_{\mathbf{M}}(X(q,S))$ can be used to separate $TR_1$ from $TR_2$: given a point $q$ from **TR**, if $E_{\mathbf{M}}(X(q,S)) \ge 1/2$, one may assign $q$ to $TR_1$, and if $E_{\mathbf{M}}(X(q,S)) < 1/2$, one may assign $q \in F$ to class 2. Since the training set **TR** is a "representative" of the feature space $F$ and $X(q,S)$ is an estimator of $E_{\mathbf{M}}(X(q,S))$, one may simply set the following classification rule: given an $S$, if $X(q,S) \ge 1/2$, assign $q \in F$ to class 1 and if $X(q,S) < 1/2$, assign $q \in F$ to class 2. The rationale in setting this classification rule is that the rule is at least reasonable for classifying the points in **TR**. Simple algebra reduces the above rule to the following. When $\mathbb{1}_S(q) = 1$, $q$ is assigned to class 1 if $r(S,TR_1) > r(S,TR_2)$ and assigned to class 2 otherwise. When $\mathbb{1}_S(q) = 0$, $q$ is assigned to class 2 if $r(S,TR_1) > r(S,TR_2)$ and assigned to class 1 otherwise. Therefore this classification rule based on $X(\cdot,S)$ is actually identical with the one given by $S$. Thus treated as one classifier, $X(\cdot,S)$ is (very) weak. However, averaging multiple weak classifiers $X(\cdot,S)$ can lead to a strong classifier.

2.3. *Classification rules.* In SD, the weak classifiers $X(\cdot,S)$ are combined via the central limit theorem. Let $\mathbf{S}^t = (S^{(1)},\ldots,S^{(t)})$ be a random sample of size $t$

from $\mathbf{M}$ (with replacement) and define, for any $q \in F$,

$$Y(q, \mathbf{S}^t) = \frac{X(q, S^{(1)}) + \cdots + X(q, S^{(t)})}{t}.$$

Given $q$, both $X(q, S)$ and $Y(q, \mathbf{S}^t)$ have the same expectation, but the variance of $Y(q, \mathbf{S}^t)$ decreases as $t$ increases. Thus $Y(q, \mathbf{S}^t)$ will work better than $X(q, S)$. Assume NUA($\varepsilon$) holds for $\mathbf{TR}$. By the central limit theorem, when $t$ is large enough and the $|\tau_i|$'s in Theorem 1 are reasonably small, there is a high probability that $Y(q, \mathbf{S}^t)$ is close to 1 for any $q$ from $TR_1$ and close to 0 for any $q$ from $TR_2$. Hence it is seen that the difference between two classes detected by $Y$ is much more obvious than that detected by $X$. Naturally one can define the following rule:

LEVEL $t$ STOCHASTIC DISCRIMINANT CLASSIFICATION RULE $R_{\mathbf{S}^t}$. For any $q \in F$, if $Y(q, \mathbf{S}^t) \geq 1/2$, classify $q$ into class 1, denoted by $R_{\mathbf{S}^t}(q) = 1$; otherwise classify $q$ into class 2, denoted by $R_{\mathbf{S}^t}(q) = 2$.

There is another aspect of the above classification rule. For a given $\mathbf{S}^t$, one can view $Y(q, \mathbf{S}^t)$ as a map from $\mathbb{R}^p$ to $\mathbb{R}^1$. Under this map, every point $q$ in the feature space $F$ becomes a real number $y$. For $i = 1, 2$, let $f_i$ denote the probability mass function of the random variable $Y(q, \mathbf{S}^t)$ for $q \in TR_i$ ($\mathbf{S}^t$ is fixed), where the uniform probability measure $P_F$ on $F$ applies. Then the original $p$-dimensional two-class problem is reduced to the univariate two-class problem where the two classes are represented by $f_1$ (class 1) and $f_2$ (class 2). Under the strict assumption of uniformity, one can show that as $t$ becomes large enough, $E_{\mathbf{M}} f_1$ is approximated by the density of a normal distribution with mean 1 and variance inversely proportional to $t$, and $E_{\mathbf{M}} f_2$ is approximated by the density of a normal distribution with mean 0 and variance inversely proportional to $t$ [see Kleinberg (1996)]. Then for any feature $y$ from the univariate two-class problem, an obvious way to classify $y$ is to allocate $y$ to class 1 if $y \geq 0.5$ and to class 2 otherwise. Naturally this leads to the above SD classification rule.

The SD classification rule simply treats $Y(q, \mathbf{S}^t)$ as a discriminant function. This rule transforms the multivariate observations $q$ to univariate observations $y$. To a certain degree, this coincides with Fisher's idea for constructing discriminant functions [Fisher (1938)].

Note that we could instead classify $q$ into class 1 iff $Y(q, \mathbf{S}^t) \geq \gamma$ for some $\gamma \in (0, 1)$ other than $\gamma = 1/2$. If misclassifying into class 1 is considered more serious than misclassifying into class 2, we may want to choose a $\gamma > 1/2$, for example. Hereafter, we pursue the simpler case with $\gamma = 1/2$.

The word "stochastic" is used here, in part for the following reason. Let us attach to each $q$ a discrete stochastic process $\{Y(q, \mathbf{S}^t),\ t = 1, 2, \ldots\}$. Suppose that the function $\varepsilon$ in Theorem 1 is small. Then for each $q \in TR_1$, the corresponding process converges (a.s.) to some value larger than $1/2$, while for each $q \in TR_2$,

the corresponding process converges (a.s.) to some value less than $1/2$. Therefore, actually the stochastic processes $\{Y(q, \mathbf{S}^t),\ t = 1, 2, \ldots\}$ are used to perform the classification task. For more information on the origin of the term "stochastic discrimination," see Kleinberg (1990).

In summary, the SD classification rule is essentially an application of resampling. One first uses resampling techniques to get a sequence of weak classifiers $S^{(1)}, S^{(2)}, \ldots, S^{(t)}$, and then employs the machinery of base random variables and the central limit theorem to combine these weak classifiers to form a strong classifier $R_{\mathbf{S}^t}$.

This procedure of combining "components" to form a strong classifier is similar to the recent work of boosting and bagging [Breiman (1996), Schapire (1990) and Freund and Schapire (1997)], but SD is quite different from these two methods. Some brief comparisons are listed below. For simplicity, we choose Real AdaBoost in Friedman, Hastie and Tibshirani (2000).

- The methods of producing the components are different.
  1. In SD, the weak classifiers $S^{(1)}, S^{(2)}, \ldots, S^{(t)}$ are independently produced from the space $\mathbf{M}$, yielding an equivalent sequence of weak classifiers $X(\cdot, S^{(1)}), X(\cdot, S^{(2)}), \ldots, X(\cdot, S^{(t)})$.
  2. In Real AdaBoost, the weak learners $f_1, f_2, \ldots, f_t$ are dependent on each other and are generated sequentially by maintaining a set of weights over the training set.
  3. Bagging uses the bootstrap samples from the training set to form the predictors $\varphi(\cdot, \mathcal{L}^{(1)}), \varphi(\cdot, \mathcal{L}^{(2)}), \ldots, \varphi(\cdot, \mathcal{L}^{(B)})$.
- The rationale underlying the combination of the components is different.
  1. SD classification is built by checking the average of $X(q, S^{(1)}), X(q, S^{(2)}), \ldots, X(q, S^{(t)})$.
  2. In Real AdaBoost, the sum of $f_1(q), f_2(q), \ldots, f_t(q)$ gives rise to estimates of the logit of the class probabilities [Friedman, Hastie and Tibshirani (2000)].
  3. The final predictor in bagging is obtained by the vote from $\varphi(q, \mathcal{L}^{(1)}), \varphi(q, \mathcal{L}^{(2)}), \ldots, \varphi(q, \mathcal{L}^{(B)})$.

**3. Performance of SD.** This section presents performance evaluation of SD classification rules. We first give some general definitions of accuracies on classification and then present some results on both training and test sets. A discussion regarding the convergence rate, overtraining-resistance and selection of parameter values is also given.

3.1. *Definitions of accuracies.* Suppose that we separate a sample $\mathbf{T}$ according to class so that we have $\mathbf{T} = \{T_1, T_2\}$, where $T_i$ represents a sample from

class $i$. Given a classification rule $R_{\mathbf{S}^t}$, define the *accuracy* of $R_{\mathbf{S}^t}$ on $\mathbf{T}$, denoted by $a(\mathbf{S}^t, \mathbf{T})$, as the proportion of feature vectors in $T_1 \cup T_2$ which are classified correctly, namely

$$a(\mathbf{S}^t, \mathbf{T}) = \frac{1}{|T_1| + |T_2|} \sum_{i=1}^{2} \sum_{q \in T_i} \mathbb{1}_{(R_{\mathbf{S}^t}(q)=i)}, \tag{6}$$

where $\mathbb{1}_{(R_{\mathbf{S}^t}(q)=i)}$ equals 1 if $R_{\mathbf{S}^t}(q) = i$ and 0 otherwise. For a fixed level $t$, the *expected accuracy* of $R_{\mathbf{S}^t}$, or the average performance of all the rules $R_{\mathbf{S}^t}$ with respect to the proportion of the correctly classified feature vectors, is defined by

$$e(t, \mathbf{T}) = E_{\mathbf{M}}(a(\mathbf{S}^t, \mathbf{T})) = \frac{1}{|T_1| + |T_2|} \sum_{i=1}^{2} \sum_{q \in T_i} E_{\mathbf{M}} \mathbb{1}_{(R_{\mathbf{S}^t}(q)=i)}. \tag{7}$$

It is seen that $a(\mathbf{S}^t, \mathbf{T})$ is the probability in the space $(F, 2^F, P_F)$ that a randomly chosen feature vector from $T_1 \cup T_2$ is classified correctly (under the classification rule $R_{\mathbf{S}^t}$), and $e(t, \mathbf{T})$ is the probability in $(F \times \mathbf{M}, 2^F \times 2^{\mathbf{M}}, P_F \times P_{\mathbf{M}})$ that a randomly chosen feature vector from $T_1 \cup T_2$ is classified correctly.

The above definitions are quite general. If $\mathbf{T} = \mathbf{TR}$, then we have $a(\mathbf{S}^t, \mathbf{TR})$ and $e(t, \mathbf{TR})$, which indicate the performance of the classifier on the training set. If $T_i$ coincides with class $i$, then $1 - a(\mathbf{S}^t, \mathbf{T})$ is simply the overall error rate of the classification rule $R_{\mathbf{S}^t}$.

In general it is difficult to find a theoretical estimate of $a(\mathbf{S}^t, \mathbf{T})$ defined in (6) by a direct computation. In this paper we will focus on estimating $e(t, \mathbf{T})$.

3.2. *Accuracies on training sets.* Note that whether a given feature point $q$ can be correctly classified depends on $E_{\mathbf{M}}(X(q, S))$. Let $\mu(q)$ denote $E_{\mathbf{M}} X(q, S)$ and let $\sigma^2(q)$ denote $\mathrm{Var}_{\mathbf{M}} X(q, S)$. We write

$$TR_1^{(1)} = \{q \in TR_1 : \mu(q) > 1/2\}, \qquad TR_2^{(1)} = \{q \in TR_2 : \mu(q) < 1/2\},$$

$$TR_1^{(2)} = \{q \in TR_1 : \mu(q) < 1/2\}, \qquad TR_2^{(2)} = \{q \in TR_2 : \mu(q) > 1/2\},$$

$$TR_1^{(3)} = \{q \in TR_1 : \mu(q) = 1/2\}, \qquad TR_2^{(3)} = \{q \in TR_2 : \mu(q) = 1/2\},$$

$$\mathbf{TR}^{(1)} = \{TR_1^{(1)}, TR_2^{(1)}\}, \qquad \mathbf{TR}^{(2)} = \{TR_1^{(2)}, TR_2^{(2)}\},$$

$$\mathbf{TR}^{(3)} = \{TR_1^{(3)}, TR_2^{(3)}\}.$$

Then $\mathbf{TR} = \mathbf{TR}^{(1)} \cup \mathbf{TR}^{(2)} \cup \mathbf{TR}^{(3)}$. If $t$ is large, it follows from the central limit theorem that $\mathbf{TR}^{(1)}$ will be correctly classified and $\mathbf{TR}^{(2)}$ will be misclassified with a high probability by $R_{\mathbf{S}^t}$, while the status of any $q$ in $\mathbf{TR}^{(3)}$ is virtually decided by flipping a fair coin. The following theorem gives some corresponding results.

THEOREM 2. *Suppose NUA*$(\varepsilon)$ *holds for* $\mathbf{M}$. *Then*

$$e(t, \mathbf{TR}^{(1)}) \geq 1 - \exp(-tB_1), \tag{8}$$

$$e(t, \mathbf{TR}^{(2)}) \leq \exp(-tB_2), \tag{9}$$

$$e(t, \mathbf{TR}^{(3)}) = \frac{1}{2} + O\left(\frac{1}{\sqrt{t}}\right), \tag{10}$$

*where the constants* $B_i = (2\beta^2\tau^{(i)})/(\beta+2)^2 > 0$ *with* $\tau^{(i)} = \inf_{q\in\mathbf{TR}^{(i)}}\{(\mu(q) - 1/2)^2\}$ *for* $i = 1, 2$.

PROOF. Recall the inequality in Hoeffding (1963): If $X_1, X_2, \ldots, X_t$ are i.i.d. and $a \leq X_i \leq b$ $(i = 1, 2, \ldots, t)$, then, with $\overline{X_t} = (X_1 + X_2 + \cdots + X_t)/t$, $P(\overline{X_t} - EX_1 \geq x) \leq \exp\{-(2tx^2)/(b-a)^2\}$ for every $x > 0$ and every positive integer $t$. This inequality will be used to prove (8) and (9). Note that for any given $q \in TR_1^{(1)} \cup TR_2^{(1)}$, $|X(q, S) - 1/2| \leq 1/\beta + 1/2$. Set $L(q, \mathbf{S}^t) = Y(q, \mathbf{S}^t) - 1/2 = \frac{1}{t}\sum_{l=1}^{t}[X(q, S^{(l)}) - 1/2]$ and $\nu(q) = E_{\mathbf{M}}(X(q, S) - 1/2)$ for $q \in TR_1^{(1)} \cup TR_2^{(1)}$. Then if $q \in TR_1^{(1)}$, $\nu(q) > 0$ and if $q \in TR_2^{(1)}$, $\nu(q) < 0$. Applying the Hoeffding inequality to the random sample $\{-(X(q, S^{(l)}) - 1/2); l = 1, 2, \ldots, t\}$, we obtain, for any fixed $q \in TR_1^{(1)}$, $P_{\mathbf{M}}(-L(q, \mathbf{S}^t) + \nu(q) \geq \nu(q)) \leq \exp(-tB_1)$. Therefore for $q \in TR_1^{(1)}$, $E_{\mathbf{M}}\mathbb{1}_{(R_{\mathbf{S}^t}(q)=1)} = P_{\mathbf{M}}(Y(q, \mathbf{S}^t) \geq 1/2) = P_{\mathbf{M}}(L(q, \mathbf{S}^t) \geq 0) \geq 1 - \exp(-tB_1)$. Again applying the Hoeffding inequality to the random sample $\{X(q, S^{(l)}) - 1/2;\ l = 1, 2, \ldots, t\}$, we have, for any fixed $q \in TR_2^{(1)}$, $P_{\mathbf{M}}(L(q, \mathbf{S}^t) \geq 0) = P_{\mathbf{M}}(L(q, \mathbf{S}^t) - \nu(q) \geq -\nu(q)) \leq \exp(-tB_1)$, and thus $E_{\mathbf{M}}\mathbb{1}_{(R_{\mathbf{S}^t}(q)=2)} = P_{\mathbf{M}}(L(q, \mathbf{S}^t) < 0) = 1 - P_{\mathbf{M}}(L(q, \mathbf{S}^t) \geq 0) \geq 1 - \exp(-tB_1)$. The proof of (8) follows from (7). The above procedure also leads to (9).

For any fixed $q \in TR_1^{(3)}$, $E_{\mathbf{M}}\mathbb{1}_{(R_{\mathbf{S}^t}(q)=1)} = P_{\mathbf{M}}(\sqrt{t}((Y(q, \mathbf{S}^t) - 1/2)/\sigma(q)) \geq 0)$. For any fixed $q \in TR_2^{(3)}$, $E_{\mathbf{M}}\mathbb{1}_{(R_{\mathbf{S}^t}(q)=2)} = P_{\mathbf{M}}(\sqrt{t}((Y(q, \mathbf{S}^t) - 1/2)/\sigma(q)) < 0)$. It follows from the Berry–Esseen theorem with Feller's bound 3 that $|E_{\mathbf{M}}\mathbb{1}_{(R_{\mathbf{S}^t}(q)=i)} - 1/2| \leq \frac{3H}{\sqrt{t}}$ for $i = 1, 2$, where $H = \sup_{q\in\mathbf{TR}^{(3)}}\{E_{\mathbf{M}}[|X(q, S) - 1/2|^3/(\sigma(q))^3]\}$. Now (7) results in (10). □

It is seen that the constants $B_i$ depend on $\beta$ and functions $\varepsilon(q)$, $\tau_1(q)$ and $\tau_2(q)$ in Theorem 1. Since $\varepsilon$ depends on $\mathbf{TR}$, $B_i$ depend on the size of the training set and the dimension of the feature space.

3.3. *Accuracies on test sets.* Theorem 2 presents a detailed look at the performance of SD on training sets. The natural question now is, How does this stochastic discrimination method perform on other data sets? To answer this

question, we need the notion of weak indiscernibility between training and test sets. Weak indiscernibility, in a certain sense, describes the fact that training and test sets are each a representative of the feature space. Suppose there is a test set $\mathbf{TE} = \{TE_1, TE_2\}$—another available set of data, where $TE_i$ is a given random sample from class $i$.

DEFINITION. **TE** is said to be $(\mathbf{M}, \delta)$ indiscernible from **TR** if for some $\delta \in [0, 1)$ and for any $S \in \mathbf{M}$, $|r(S, TR_i) - r(S, TE_i)| \leq \delta$ for $i = 1, 2$.

This $(\mathbf{M}, \delta)$ indiscernibility is also referred to as *weak indiscernibility*.

Note that $r(S, TR_i)$ and $r(S, TE_i)$ are just two empirical estimates of the probability that a random vector $q$ from population $i$ (class $i$) falls into $S$. Thus if the sample sizes of the $TR_i$'s and the $TE_i$'s are large or the "volume" of each $S$ is big (e.g., when $\lambda$ and $a$ are large), then **TE** should be $(\mathbf{M}, \delta)$ indiscernible from **TR** for some small $\delta$. In general, such a $\delta$ largely depends on the sizes of **TR** and **TE**, $\lambda$ and $a$.

Indiscernibility in Kleinberg (1996) is stronger than the above weak indiscernibility. In Kleinberg (1996), a set $\mathbf{M}$ of $2^F$ which makes **TR** indiscernible from **TE** satisfies the condition that $r(S, TR_i) = r(S, TE_i)$ for any $S$ in **TR**. It is difficult to find such an $\mathbf{M}$, built from **TR**.

With the notion of weak indiscernibility, we will be able to see why SD works well for test data. Note that our entire development so far concerning the training data can be carried out for the test data. Let us begin with the near uniformity assumption. Note that partition (2) is made according to $TR_1$ and $TR_2$. Now we partition $\mathbf{M}$ in terms of $\{TE_1, TE_2\}$ and assume the corresponding near uniformity assumption with the involved function $\varepsilon^*(q)$, denoted by NUA$^*(\varepsilon^*)$. For convenience, quantities involving test sets will be indicated with an asterisk $(*)$ flag. Our first result concerning the test data set is the following theorem, a counterpart of Theorem 1.

THEOREM 3. *Suppose that there exists a* $\delta$ $(< \beta/2)$ *for which* **TR** *is* $(\mathbf{M}, \delta)$ *indiscernible from* **TE**. *Assume NUA*$^*(\varepsilon^*)$ *holds for* $\mathbf{M}$. (a) $E_{\mathbf{M}}(X(q, S)) = 1 + \tau_1^*(q) + \alpha(q)$ *if* $q \in TE_1$ *and* $= \tau_2^*(q) + \alpha(q)$ *if* $q \in TE_2$; (b) $\mathrm{Var}_{\mathbf{M}}(X(q, S)) \leq [(v(q))^{1/2} + 4\delta/(\beta(\beta - 2\delta))]^2$, *where* $|\tau^*| \leq \varepsilon^*(q)/(\beta - 2\delta)$, $|\alpha(q)| \leq 4\delta/(\beta(\beta - 2\delta))$ *and* $v(q) \leq (1 + 4\varepsilon^*(q))/(4(\beta - 2\delta)^2) + 1$.

PROOF. Corresponding to the base random variable $X(q, S)$ defined in (1), a random variable based on $r(S, TE_i)$ can be studied. Since there exists a $\delta$ $(< \beta/2)$ for which **TR** is $(\mathbf{M}, \delta)$ indiscernible from **TE**, it is seen that $|r(S, TE_1) - r(S, TE_2)| \geq \beta - 2\delta > 0$. For any $(q, S) \in F \times \mathbf{M}$, define

$$X^*(q, S) = \frac{\mathbb{1}_S(q) - r(S, TE_2)}{r(S, TE_1) - r(S, TE_2)}.$$

By NUA*($\varepsilon^*$), we see as in Theorem 1 that $E_{\mathbf{M}}(X^*(q,S)) = 1+\tau_1^*(q)$ if $q \in TE_1$ and $= \tau_2^*(q)$ if $q \in TE_2$, and $\mathrm{Var}_{\mathbf{M}}(X^*(q,S)) \le (1+4\varepsilon^*(q))/(4(\beta-2\delta)^2) - (\tau_1^*(q))^2 - 2\tau_1^*(q)$ if $q \in TE_1$ and $\le (1+4\varepsilon^*(q))/(4(\beta-2\delta)^2) - (\tau_2^*(q))^2$ if $q \in TE_2$, where $|\tau^*| \le \varepsilon^*(q)/(\beta-2\delta)$.

From the definitions of $X$ and $X^*$, one can show that for any $(q,S) \in F \times \mathbf{M}$, $|X^*(q,S) - X(q,S)| \le 4\delta/(\beta(\beta-2\delta))$. Since $|E_{\mathbf{M}}(X(q,S) - X^*(q,S))| \le E_{\mathbf{M}}|X(q,S) - X^*(q,S)|$, we have $E_{\mathbf{M}}(X(q,S)) = E_{\mathbf{M}}(X^*(q,S)) + \alpha(q) = 1 + \tau_1^*(q) + \alpha(q)$ if $q \in TE_1$ and $= \tau_2^*(q) + \alpha(q)$ if $q \in TE_2$, where $|\alpha(q)| \le 4\delta/(\beta(\beta-2\delta))$. Also $\mathrm{Var}_{\mathbf{M}}(X) = \mathrm{Var}_{\mathbf{M}}(X-X^*) + \mathrm{Var}_{\mathbf{M}}(X^*) + 2\,\mathrm{cov}(X - X^*, X^*) \le \mathrm{Var}_{\mathbf{M}}(X-X^*) + \mathrm{Var}_{\mathbf{M}}(X^*) + 2(\mathrm{Var}_{\mathbf{M}}(X-X^*))^{1/2}(\mathrm{Var}_{\mathbf{M}}(X^*))^{1/2} \le (4\delta/(\beta(\beta-2\delta)))^2 + v(q) + 8\delta/(\beta(\beta-2\delta))(v(q))^{1/2}$, where $v(q) = \mathrm{Var}_{\mathbf{M}}(X^*) \le (1+4\varepsilon^*(q))/(4(\beta-2\delta)^2) + 1$. □

Recall that the SD classification rule works for the training data set **TR** simply because $E_{\mathbf{M}}(X(q,S))$ separates $TR_1$ from $TR_2$ (see the discussion following Theorem 1). Theorem 3 shows that when $\tau_1^*$, $\tau_2^*$ and $\alpha$ are small, $E_{\mathbf{M}}(X(q,S))$ is close to 1 for $q \in TE_1$ and close to 0 for $q \in TE_2$, that is, $E_{\mathbf{M}}(X(q,S))$ also separates $TE_1$ from $TE_2$. Therefore under the assumption of NUA*($\varepsilon^*$) and $(\mathbf{M},\delta)$ indiscernibility ($\delta < \beta/2$) between training and test sets, the rationale that the stochastic discriminant classification rule $R_{\mathbf{S}^t}$ works for the test set **TE** is exactly the same as for the training set **TR**. Hence it can be expected that the performance of SD on the training set will be projected on the test set. An immediate consequence of this projectability is that results similar to Theorem 2 hold for **TE**.

Let

$$TE_1^{(1)} = \{q \in TE_1 : \mu(q) > 1/2\}, \qquad TE_2^{(1)} = \{q \in TE_2 : \mu(q) < 1/2\},$$
$$TE_1^{(2)} = \{q \in TE_1 : \mu(q) < 1/2\}, \qquad TE_2^{(2)} = \{q \in TE_2 : \mu(q) > 1/2\},$$
$$TE_1^{(3)} = \{q \in TE_1 : \mu(q) = 1/2\}, \qquad TE_2^{(3)} = \{q \in TE_2 : \mu(q) = 1/2\},$$
$$\mathbf{TE}^{(1)} = \{TE_1^{(1)}, TE_2^{(1)}\}, \qquad \mathbf{TE}^{(2)} = \{TE_1^{(2)}, TE_2^{(2)}\},$$
$$\mathbf{TE}^{(3)} = \{TE_1^{(3)}, TE_2^{(3)}\}.$$

We have the following:

THEOREM 4. *Suppose that* NUA*($\varepsilon^*$) *holds for* $\mathbf{M}$ *and that there exists a* $\delta < \beta/2$ *for which* **TR** *is* $(\mathbf{M},\delta)$ *indiscernible from* **TE**. *Then*

$$e(t, \mathbf{TE}^{(1)}) \ge 1 - \exp(-tB_1^*), \tag{11}$$

$$e(t, \mathbf{TE}^{(2)}) \le \exp(-tB_2^*), \tag{12}$$

$$e(t, \mathbf{TE}^{(3)}) = \frac{1}{2} + O\left(\frac{1}{\sqrt{t}}\right), \tag{13}$$

*where the constants* $B_i^*$ $(> 0)$ *depend on* $\beta$, $\delta$ *and functions* $\alpha(q)$, $\varepsilon^*(q)$, $\tau_1^*(q)$ *and* $\tau_2^*(q)$.

PROOF. From Theorem 2 or its proof, the statements hold. □

Note that the constants $B_i^*$ depend on the dimension of the feature space and the sizes of the training and test sets. This is simply because $\varepsilon^*$ is dependent on **TE** and thus the dimension of the feature space, and $\delta$ is dependent on the sizes of **TR** and **TE**.

3.4. *Convergence rate, overtraining-resistance and parameter tuning.* In this section, we discuss the convergence rate, overtraining-resistance and selection of parameters.

3.4.1. *Convergence rate.* Sections 3.2 and 3.3 break **TR** and **TE** down into $\mathbf{TR}^{(i)}$ and $\mathbf{TE}^{(i)}$: $\mathbf{TR} = \mathbf{TR}^{(1)} \cup \mathbf{TR}^{(2)} \cup \mathbf{TR}^{(3)}$ and $\mathbf{TE} = \mathbf{TE}^{(1)} \cup \mathbf{TE}^{(2)} \cup \mathbf{TE}^{(3)}$. We conjecture that both $r(\mathbf{TR}^{(3)}, \mathbf{TR})$ and $r(\mathbf{TE}^{(3)}, \mathbf{TE})$ are negligible. This simply states that each of $\mathbf{TR}^{(3)}$ and $\mathbf{TE}^{(3)}$ has a very small size compared with **TR** and **TE**, respectively. Neglecting the size of $\mathbf{TR}^{(3)}$ and $\mathbf{TE}^{(3)}$, one sees, from (8), (9), (11) and (12), that both $e(t, \mathbf{TR})$ and $e(t, \mathbf{TE})$ converge at least exponentially fast. This indicates that SD is a fast algorithm with respect to $t$.

3.4.2. *Overtraining-resistance.* Overtraining-resistance is one of the important properties of SD. The empirical evidence is that good performance of SD on training data translates into good performance on test data. In theory, we believe that overtraining is prevented by the weak indiscernibility between training and test data sets, which is usually controlled by $\lambda$, $a$ and $b$. A supporting argument for a special case is given as follows. Suppose $\varepsilon \approx 0$. Then from Theorem 1, $\mu(q) \approx 1$ for $q \in TR_1$ and $\mu(q) \approx 0$ for $q \in TR_2$. From Theorem 2, we see that **TR** is perfectly classified by SD. When will **TE** be perfectly classified? To answer this question, we assume that the natural condition $\varepsilon^* \approx 0$ holds and that there exists $(\mathbf{M}, \delta)$ indiscernibility between **TR** and **TE** $(\delta < \beta/2)$. Then from Theorem 3, $\mu(q) \approx 1 + \alpha(q)$ for $q \in TE_1$ and $\mu(q) \approx \alpha(q)$ for $q \in TE_2$. Since the magnitude of $\alpha$ is unknown, one cannot be sure that **TE** will be perfectly classified. However, if we further assume that $\delta \approx 0$, then $\alpha(q) \approx 0$ and Theorem 4 indicates that **TE** is perfectly classified.

3.4.3. *Tuning parameters.* It appears that parameters $\lambda$, $a$, $b$ and $\beta$ have a strong influence on the performance of SD.

In fact, it is seen that $\lambda$, $a$ and $b$ determine the volume of $S$. A big volume is required for weak indiscernibility to hold so that good performance of SD on training data can translate into good performance on test data.

Theoretically, large $\beta$ is required. There are several reasons for this. First, let us note that the size of $\mathbf{TR}^{(2)} \cup \mathbf{TR}^{(3)}$ and the size of $\mathbf{TE}^{(2)} \cup \mathbf{TE}^{(3)}$ contribute to the training and test error rates, respectively. From the definitions of $\mathbf{TR}^{(2)}$, $\mathbf{TR}^{(3)}$, $\mathbf{TE}^{(2)}$ and $\mathbf{TE}^{(3)}$, the sizes of these sets can be reduced by choosing large $\beta$ and small $\delta$, since there will be more points $q$ in $TR_1$ and $TE_1$ such that $\mu(q) > 1/2$ and more points $q$ in $TR_2$ and $TE_2$ such that $\mu(q) < 1/2$. Second, the condition $\delta < \beta/2$ is more likely satisfied when large $\beta$ is used. Third, the upper bound of $\tau$ in Theorem 1 may become smaller when $\beta$ is larger and, consequently, $E_{\mathbf{M}}(X(q, S))$ for $q \in TR_1$ is closer to 1 and $E_{\mathbf{M}}(X(q, S))$ for $q \in TR_2$ is closer to 0. This shows that with larger $\beta$, weak classifiers may "discern" the two classes more easily. A disadvantage of using larger $\beta$ is that the training process usually takes more time.

Since the quantitative relationship among these parameters is unavailable, we can only obtain a near optimal result by selecting an appropriate combination of $\lambda$, $a$, $b$ and $\beta$. As discussed above, $\lambda$, $a$ and $b$ are used to determine the volume of a weak classifier. For convenience, we fix $a = 0.1$ and $b = 1.0$. Then tuning $\lambda$ and $\beta$ can be done through cross-validation or the usual training/test procedure. We simply run SD by stepping through some ranges of $\lambda$ and $\beta$ to find out the appropriate values for these two parameters that correspond to the best test performance achieved. The range of $\lambda$ may be set to be $\lambda \in [1, 2]$ and the range of $\beta$ may be set to be $\beta \in [0.05, 0.95]$. Examples of this tuning process are seen in Section 4.

**4. Experimental results.** In this section, we report some results of our experiments conducted on one simulated data set and two popular data sets from the repository at the University of California at Irvine. The experiments are used to provide a simple look at how SD works in practice.

In all the experiments, only one set of values of $\lambda$, $a$, $b$ and $\beta$ was used for all the runs associated with each data set. The selection of the parameters was made in the following way (see Section 3.4.3). We used $a = 0.1$ and $b = 1.0$ for all the data sets. Tuning $\lambda$ and $\beta$ was done through fivefold cross-validation for Examples 1 and 2, and via the usual training/test procedure for the simulated data in Example 3. This tuning process consisted of two steps. In step 1, we conduced a coarse tuning. We considered $\lambda \in [1.0, 2.0]$ and $\beta \in [0.05, 0.95]$. We ran SD for each choice of $\lambda$ and $\beta$ by looping over the ranges with a step size of 0.1 for both $\lambda$ and $\beta$. For Examples 1 and 2, we selected the combination of $\lambda$ and $\beta$ that corresponded to the best averaged test performance. For Example 3, we fixed a training set of size 400 and a test set of size 4000, and then we selected the combination of $\lambda$ and $\beta$ that corresponded to the best performance on the test data set. Denote the selected parameters by $\lambda_0$ and $\beta_0$. In step 2, we conducted a fine tuning. We considered new ranges of $\lambda$ and $\beta$ centered at $\lambda_0$ and $\beta_0$, respectively. The length of each range was set to be half of that in step 1. An obvious truncation was done if the new range extended beyond the range in step 1. We ran SD for

each choice of $\lambda$ and $\beta$ by looping over the ranges with a step size of 0.05 for both $\lambda$ and $\beta$. As in step 1, we chose the combination of $\lambda$ and $\beta$ that corresponded to the best test performance as the fine tuning result. Denote the selected parameter values by $\lambda_1$ and $\beta_1$. These values $\lambda_1$ and $\beta_1$ were used with the actual runs of the experiments. (During the tuning process, a combination of $\lambda$ and $\beta$ was discarded if, with this combination, it took a long time to locate a weak classifier.)

EXAMPLE 1 (Breast cancer, Wisconsin). The data came from Dr. William H. Wolberg, University of Wisconsin Hospitals, Madison [Wolberg and Mangasarian (1990)]. The data set contains 699 points in the nine-dimensional space $\mathbb{R}^9$ that come from two classes: benign (458 cases) and malignant (241 cases). We used fivefold cross-validation to estimate the test error rate and we reran each cross-validation 10 times using different seeds. The SD test error is 3.1%. Breiman (1996) reported an error rate of 3.7% from bagging. Friedman, Hastie and Tibshirani (2000) reported various error rates from different versions of boosting: their three lowest error rates corresponding to 200 iterations were 2.9, 3.1 and 3.2%.

EXAMPLE 2 (Pima Indians, diabetes). The data were gathered among the Pima Indians by the National Institute of Diabetes and Digestive and Kidney Disease [Smith, Everhart, Dickson, Knowler and Johannes (1988)]. The data set contains 768 points in the space $\mathbb{R}^8$ from two classes: tested positive (268 cases) or tested negative (500 cases). We used fivefold cross-validation to estimate the test error rate and we reran each cross-validation 10 times using different seeds. The test error of SD is 26.2%. The five lowest error rates from bagging and boosting reported in Freund and Schapire (1996) are 24.4, 25.3, 25.7, 26.1 and 26.4%, where tenfold cross-validation was used.

EXAMPLE 3 (Two classes with the same mean). Consider two ten-dimensional normal distributions $N(\mathbf{0}, \mathbf{I})$ and $N(\mathbf{0}, 1.85\mathbf{I})$, where $\mathbf{I}$ is the $10 \times 10$ identity matrix. Let $\pi_1 = \pi_2 = 1/2$ so that the Bayes error is 25%. The training set $\mathbf{TR}$ contains 400 points from each class. This is the example used in Friedman, Hastie and Tibshirani (2000) for an illustration of overfitting of boosting. The averaged results over 10 independently drawn training/test set combinations were used to estimate the error rates. Overfitting does not occur here. Figure 3 shows the performance of SD.

The above examples were simply used to demonstrate the application of the SD algorithm in Figure 2. We have not tried to find the best setting of the parameters $\lambda$, $a$, $b$ and $\beta$. Although the tuning process proposed above often works well, it is in no way the best or the unique method. As better methods are found to pick up the parameters, the classification results will definitely be improved. For more experimental results of SD, we refer the readers to Kleinberg (2000), where a technique called uniformity forcing was incorporated

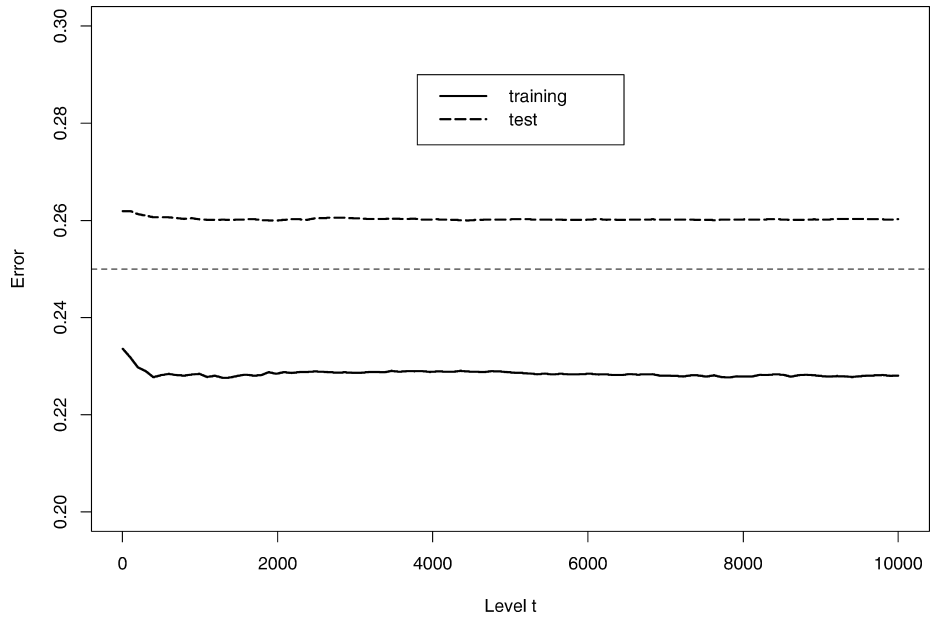


FIG. 3. *Training and test errors for two ten-dimensional normal distributions with the same mean. The training set contains* 400 *points from each class. The dashed horizontal line indicates the Bayes error* 25%.

into the algorithmic implementation of SD. There the performance of SD on 24 public data sets was compared with those of various pattern recognition methods, including boosting and bagging, and the results showed that SD placed first on 19 of them, second on two others, fourth on another, and fifth on the remaining two.

**5. Conclusion.** SD, treated as a statistical method in pattern recognition, does not fall into the classical sampling paradigm or the diagnostic paradigm. [For the definitions of sampling and diagnostic paradigms, see Ripley (1996), page 27.] This paper studies SD for two-class classification under relaxed assumptions. Uniformity and indiscernibility have been replaced by near uniformity and weak indiscernibility, respectively. An algorithm implementing SD is described. In addition, theoretical results of the classification accuracy on both training and test sets are provided to judge the performance of SD. In practice, SD, a method especially suitable to parallel implementation, is effective.

The two-class case is the core of the whole structure of SD. Higher class pattern recognition using SD may be realized on the basis of two-class classification, and the major results developed in this paper can be extended to the multiclass situation.

**Acknowledgments.** The first author expresses deep gratitude to his advisor, Professor Eugene Kleinberg, for introducing the field of pattern recognition and sharing his own research. We thank the Editors, Associate Editors and referees for their thorough review of earlier versions and very helpful comments that led to substantial improvements.

## REFERENCES

BERLIND, R. (1994). An alternative method of stochastic discrimination with applications to pattern recognition. Ph.D. dissertation, Dept. Mathematics, State Univ. New York, Buffalo.

BREIMAN, L. (1996). Bagging predictors. *Machine Learning* **24** 123–140.

CHEN, D. (1998). Estimates of classification accuracies for Kleinberg's method of stochastic discrimination in pattern recognition. Ph.D. dissertation, Dept. Mathematics, State Univ. New York, Buffalo.

DUDA, R. O., HART, P. E. and STORK, D. G. (2001). *Pattern Classification*, 2nd ed. Wiley, New York.

FISHER, R. A. (1938). The statistical utilization of multiple measurements. *Annals of Eugenics* **8** 376–386.

FREUND, Y. and SCHAPIRE, R. E. (1996). Experiments with a new boosting algorithm. In *Proc. 13th International Conference on Machine Learning* 148–156. Morgan Kauffman, San Francisco.

FREUND, Y. and SCHAPIRE, R. E. (1997). A decision–theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.* **55** 119–139.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion). *Ann. Statist.* **28** 337–407.

FUKUNAGA, K. (1990). *Introduction to Statistical Pattern Recognition*, 2nd ed. Academic Press, New York.

HO, T. K. (1995). Random decision forests. In *Proc. Third International Conference on Document Analysis and Recognition* (M. Kavanaugh and P. Storms, eds.) **1** 278–282. IEEE Computer Society Press, New York.

HO, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20** 832–844.

HO, T. K. and BAIRD, H. S. (1998). Pattern classification with compact distribution maps. *Computer Vision and Image Understanding* **70** 101–110.

HO, T. K. and KLEINBERG, E. M. (1996). Building projectable classifiers of arbitrary complexity. In *Proc. 13th International Conference on Pattern Recognition* (M. E. Kavanaugh and B. Werner, eds.) **2** 880–885. IEEE Computer Society Press, New York.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

KLEINBERG, E. M. (1990). Stochastic discrimination. *Ann. Math. Artif. Intell.* **1** 207–239.

KLEINBERG, E. M. (1996). An overtraining-resistant stochastic modeling method for pattern recognition. *Ann. Statist.* **24** 2319–2349.

KLEINBERG, E. M. (2000). On the algorithmic implementation of stochastic discrimination. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22** 473–490.

KLEINBERG, E. M. and HO, T. K. (1993). Pattern recognition by stochastic modeling. In *Proc. Third International Workshop on Frontiers in Handwriting Recognition* (M. Bosker and R. Casey, eds.) 175–183. Partners Press, Buffalo.

MCLACHLAN, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, New York.

RIPLEY, B. D. (1996). *Pattern Recognition and Neural Networks.* Cambridge Univ. Press.

SCHAPIRE, R. E. (1990). The strength of weak learnability. *Machine Learning* **5** 197–227.

SMITH, J., EVERHART, J., DICKSON, W., KNOWLER, W. and JOHANNES, R. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proc. Symposium on Computer Applications and Medical Care* 261–265. IEEE Computer Society Press, New York.

WOLBERG, W. H. and MANGASARIAN, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proc. Nat. Acad. Sci. U.S.A.* **87** 9193–9196.

D. CHEN
DEPARTMENT OF PREVENTIVE MEDICINE
AND BIOMETRICS
UNIFORMED SERVICES UNIVERSITY
OF THE HEALTH SCIENCES
BETHESDA, MARYLAND 20814
E-MAIL: dchen@usuhs.mil

P. HUANG
DEPARTMENT OF BIOMETRY
AND EPIDEMIOLOGY
MEDICAL UNIVERSITY OF SOUTH CAROLINA
CHARLESTON, SOUTH CAROLINA 29425
E-MAIL: huangp@musc.edu

X. CHENG
GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC 20052
E-MAIL: cheng@gwu.edu