

Short Paper

On Kleinberg's Stochastic Discrimination Procedure

Albrecht Irle and Jonas Kauschke

Abstract—A new condition for high accuracy on test sets is given for the method of stochastic discrimination (SD), a pattern recognition method introduced by Kleinberg. This condition provides a simple explanation for the observed good generalization properties and overtraining-resistance. We also show that the method of SD remains valid if the original assumption of uniform distribution on a finite space for resampling is relaxed.

Index Terms—Pattern recognition, stochastic discrimination.

1 INTRODUCTION

AN ensemble method for pattern recognition using resampling of classifiers was devised and investigated by Kleinberg, see [10], [11]. This method of stochastic discrimination (SD) starts from a space of weak classifiers. These are combined by resampling and averaged to form strong classifiers with largely improved performance. This method was shown to have the important properties of high accuracy on training and test set and overtraining-resistance. The conditions of [11] for achieving high accuracy, i.e., small error rates on the training set and test set, were weakened in [2]. Algorithms for generating weak classifiers in practical implementations were provided in [12], [2]. These papers also report on the performance in various synthetic and real data problems and their findings indicate that SD is a very promising method for pattern recognition. Further research on this method can, e.g., be found in [8].

Although SD was introduced already in 1990 in [10], it has not yet found textbook coverage as other popular ensemble methods such as bagging and boosting have. So it may be useful to include a short discussion of similarities and differences of these methods and SD, which is done in Section 7. For a thorough textbook treatment of bagging and boosting, we refer to [7, 8.7 and 10.1 to 10.6].

The main purpose of this note is the presentation of a new condition for high accuracy on test sets in Section 5. This condition gives a simple explanation for the good generalization property of SD, i.e., that a good performance on the training data leads to a comparable performance on the test data, and for the observed overtraining-resistance.

SD has brought two major new ideas to pattern recognition. The first is given by the idea of resampling from a set $\mathcal{H}$ of weak classifiers to form a strong classifier. The second may be described as giving general principles for the construction of weak classifiers which form strong classifiers when resampled.

In Sections 2-4, we review the results of [11], [2] in a somewhat generalized framework which shows that the original assumptions of the finiteness of the space of observables and of the space of classifiers and the assumption of strictly uniform resampling

are not essential. Resampling is treated in Section 2, the construction of weak classifiers in Sections 3-4. Section 6 gives an illustrative example for the performance of SD. In our notation, we follow [11], [2].

2 A RESAMPLING METHOD FOR COMBINING CLASSIFIERS

Let F denote the sample space of the observables q ; the unobservables may take only the two values 1, 2 so we consider a two class discrimination problem with classes 1 and 2.

Let us call any bounded mapping

$$h : F \rightarrow \mathbb{R},$$

a classification function, c-function for short, with corresponding classifier

$$c_h : F \rightarrow \{1, 2\}, \quad c_h(q) = \begin{cases} 1 & h(q) \geq 0 \\ 2 & h(q) < 0 \end{cases}.$$

Let $\mathcal{H}$ be a set of c-functions from which we resample. Fix a probability measure $P_{\mathcal{H}}$ on $\mathcal{H}$, where $\mathcal{H}$ is supplied with a suitable σ -algebra such that the mappings $h \mapsto h(q)$ are measurable for all $q \in F$. $P_{\mathcal{H}}$ provides the distribution for resampling c-functions.

Resampling by independent draws with replacement from a finite set of observed data are commonly known as the bootstrap in statistics, leading to a sequence of independent, identically distributed random variables. For SD, we do not resample from some observed data but do the resampling from the space $\mathcal{H}$ of classification functions by drawing independently at each step according to the distribution $P_{\mathcal{H}}$.

This resampling is formally described by an independent, identically distributed sequence $H_1, H_2, \dots$ of random variables taking values in $\mathcal{H}$. They are defined on some probability space with underlying probability measure P and the distribution of each H_i is $P_{\mathcal{H}}$, i.e., for any measurable $A \subseteq \mathcal{H}$,

$$P(H_i \in A) = P_{\mathcal{H}}(A).$$

So, H_i is the c-function resulting from the i th resampling step. Having generated $H_1, H_2, \dots, H_n$, we then use the c-function:

$$H^{(n)} = \frac{1}{n} \sum_{i=1}^n H_i.$$

The classifier becomes

$$c_{H^{(n)}}(q) = \begin{cases} 1 & H^{(n)}(q) \geq 0 \\ 2 & H^{(n)}(q) < 0 \end{cases}.$$

For a given $q \in F$, the probabilities that the randomly picked $c_{H^{(n)}}$ misclassifies q , i.e., the probabilities of misclassification due to resampling, are given by

$$P\left(\frac{1}{n} \sum_{i=1}^n H_i(q) \leq 0\right) = P(c_{H^{(n)}}(q) = 2)$$

for q belonging to class 1,

$$P\left(\frac{1}{n} \sum_{i=1}^n H_i(q) > 0\right) = P(c_{H^{(n)}}(q) = 1)$$

for q belonging to class 2.

Note that probabilities and expectations are always computed with respect to the resampling distribution, and that no probabilistic model for the observables and classes is involved.

- The authors are with the Mathematical Institute, University of Kiel, Ludewig-Meyn-Straße 4, D-24098 Kiel, Germany.
E-mail: {irle, kauschke}@math.uni-kiel.de.

Manuscript received 10 Feb. 2010; revised 27 July 2010; accepted 28 July 2010; published online 13 Dec. 2010.

Recommended for acceptance by K. Murphy.

For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2010-02-0089.

Digital Object Identifier no. 10.1109/TPAMI.2010.225.

Assume now that for a fixed $(q, j) \in F \times \{1, 2\}$ with class $j = 1$, the resampling distribution has the property that the randomly drawn weak classifier fulfills $EH_1(q) > 0$ and hence shows the right tendency toward classifying this q .

By large deviation theory, assuming finiteness of the moment generating function, this implies that the *probability of misclassification due to resampling converges to zero exponentially fast*. The same exponential convergence to zero holds for $(q, j) \in F \times \{1, 2\}$ with class $j = 2$ if $EH_1(q) < 0$. The precise statement is given in the next lemma, which uses the moment generating function ϕ of $H_1(q)$.

Lemma 2.1. *Let $(q, j) \in F \times \{1, 2\}$ and assume that $\phi(t) = E \exp(tH_1(q)) < \infty$ for t in some neighborhood of zero. Then, $\phi^* = \sup_t [-\log \phi(t)] > 0$ and*

$$1. \quad \text{for } j = 1 \text{ and } EH_1(q) > 0,$$

$$P(c_{H^n(q)} = 2) \leq e^{-n\phi^*} \text{ for all } n;$$

$$2. \quad \text{for } j = 2 \text{ and } EH_1(q) < 0,$$

$$P(c_{H^n(q)} = 1) \leq e^{-n\phi^*} \text{ for all } n.$$

Proof. This restates the standard upper exponential estimate of large deviation theory, see e.g., [3, 1.9]. For completeness, we sketch the proof of item 2. Since $\phi(0) = 1$, $\phi'(0) = EH_1(q) < 0$, and ϕ is strictly convex on $\{t : \phi(t) < \infty\}$, we obtain $\inf_t \phi(t) = \inf_{t>0} \phi(t) < 1$ and $\phi^* > 0$. For any $t > 0$ we have

$$\begin{aligned} P(c_{H^n(q)} = 1) &= P\left(\sum_{i=1}^n H_i(q) > 0\right) \\ &= P\left(\exp\left(t \sum_{i=1}^n H_i(q)\right) > 1\right) \\ &\leq E \exp\left(t \sum_{i=1}^n H_i(q)\right) = \phi(t)^n, \end{aligned}$$

which implies item 2. $\square$

This key observation of exponentially small probabilities of misclassification as a consequence of resampling classifiers is due to Kleinberg, see [11], also [2, Theorems 2 and 4]. In these references, exponential bounds are given for uniformly bounded $\mathcal{H}$ applying Hoeffding's inequality. The consequences for classification theory may be phrased in the following terms, see [11]. Call $(\mathcal{H}, P_{\mathcal{H}})$ a weak model for (q, j) if

$$\begin{aligned} EH_1(q) &> 0, \text{ for class } j = 1, \\ EH_1(q) &< 0, \text{ for class } j = 2. \end{aligned}$$

Generating a c-function h from such a weak model for (q, j) yields a classifier which may be called a weak learner in the following sense: The probability, with respect to $P_{\mathcal{H}}$, of producing a wrong classification for (q, j) may be large, although due to the condition on the expectation, there is a tendency of producing the correct classification.

Combining weak models by resampling from $P_{\mathcal{H}}$ leads to the classifier $c_{H^{(n)}}$ which has an exponentially small probability in n of misclassifying (q, j) . As is common in resampling procedures, the number n is at our disposal and is only limited by time and computational constraints. So by generating a large number of weak learners, we can produce a strong learner which has high accuracy for all (q, j) in the training set as long as A2 holds. In practice, there might be some elements in the training sequence violating A2 and, for those, no conclusion on the quality of their classification seems possible.

Let us remark that this convergence of misclassification probabilities to zero due to resampling is not related to any

convergence of the Bayes error if we had, in addition, a probabilistic model for the observables and classes.

3 THE STOCHASTIC DISCRIMINATION PROCEDURE AND ITS BEHAVIOR ON THE TRAINING SET

Let us assume now that we are given a finite training set $TR \subseteq F \times \{1, 2\}$ in the form

$$TR = TR_1 \times \{1\} \cup TR_2 \times \{2\},$$

where $TR_i = \{q \in F : (q, i) \in TR\}$, $i = 1, 2$, and we assume that TR_1 and TR_2 are disjoint subsets of F . In [11], [2], c-functions of the form

$$a(S)1_S - b(S)$$

are considered. Here, 1_S is the indicator function of a set $S \in \mathcal{M}$, where $\mathcal{M}$ is a certain system of subsets of F and $a(S), b(S)$ are normalizing constants; hence,

$$\mathcal{H} = \{a(S)1_S - b(S) : S \in \mathcal{M}\}.$$

The probability measure $P_{\mathcal{H}}$ now becomes a probability measure $P_{\mathcal{M}}$ on $\mathcal{M}$, supplied with a suitable σ -algebra such that the mappings $S \mapsto 1_S(q)$ are measurable for all $q \in F$. So, resampling of sets in $\mathcal{M}$ is done with respect to $P_{\mathcal{M}}$. Our review will show that the assumptions in [11], [2] that F and $\mathcal{M}$ are finite and that $P_{\mathcal{M}}$ is the uniform distribution on $\mathcal{M}$ are not essential for their results as long as the important near uniformity assumption A2, stated below, is not violated.

In [11], conditions were found such that $\mathcal{M}$ provides weak learners for all $(q, j) \in TR$ in the sense of Section 2. These conditions provide the proper insight but would be met only approximately in practice, and weakened conditions were given in [2]. Let us now state these conditions. For shorthand notation, we write $r(A|B) = \frac{|A \cap B|}{|B|}$ for $A, B \subseteq F$.

A1.

$$\min_{S \in \mathcal{M}} |r(S|TR_1) - r(S|TR_2)| = \beta > 0.$$

Note that the set

$$\{(r(S|TR_1), r(S|TR_2)) : S \in \mathcal{M}\}$$

is finite due to the finiteness of $TR_1 \cup TR_2$. We write this set as $\{x^{(1)}, \dots, x^{(k)}\}$, with $x^{(l)} = (x_1^{(l)}, x_2^{(l)})$, and set

$$\mathcal{M}_l = \{S \in \mathcal{M} : (r(S|TR_1), r(S|TR_2)) = x^{(l)}\},$$

so that $\mathcal{M}_1, \dots, \mathcal{M}_k$ forms a partition of $\mathcal{M}$.

For the following, let $\gamma \geq 0$.

A2. Near Uniformity Assumption $NUA(\gamma, TR)$, see [2].

There exists a function $\varepsilon : TR_1 \cup TR_2 \rightarrow [0, \gamma]$ such that, for all $l = 1, \dots, k$, $i = 1, 2$, and $q \in TR_i$,

$$|P_{\mathcal{M}}(q \in S|\mathcal{M}_l) - x_i^{(l)}| \leq \varepsilon(q).$$

The uniformity assumption of [11] requires $\varepsilon(q) = 0$. The following observation, which holds for arbitrary probability measures $P_{\mathcal{M}}$, illustrates this assumption.

Remark. For all $l = 1, \dots, k$, $i = 1, 2$,

$$x_i^{(l)} = \frac{1}{|TR_i|} \sum_{q \in TR_i} P_{\mathcal{M}}(q \in S|\mathcal{M}_l).$$

This is immediate from

$$\begin{aligned} \sum_{q \in TR_i} P_{\mathcal{M}}(q \in S | \mathcal{M}_i) &= \sum_{q \in TR_i} E_{\mathcal{M}}(1_S(q) | \mathcal{M}_i) \\ &= E_{\mathcal{M}} \left(\sum_{q \in TR_i} 1_S(q) | \mathcal{M}_i \right) = E_{\mathcal{M}}(S \cap TR_i | \mathcal{M}_i) \\ &= x_i^{(i)} | TR_i | \text{ by definition of } \mathcal{M}_i. \end{aligned}$$

So, A2 means that by resampling from any $\mathcal{M}_i$, the elements of TR_i are treated equally up to a difference of order at most γ . Resampling from $\mathcal{M}$ may be viewed in the following way: First, draw one of the finitely many $\mathcal{M}_i$ randomly and then draw one of the elements of that $\mathcal{M}_i^*$ which resulted in the first draw. The second draw is addressed by A2 and should be close to uniform in order to confirm to A2, whereas the first draw of picking some $\mathcal{M}_i$ has no influence on the validity of A2 and may be chosen in a general way.

We now turn to the *definition of the c-functions*, [11], [2]. For $S \in \mathcal{M}$, define the c-functions

$$X_S = 2 \left(\frac{1_S - r(S|TR_2)}{r(S|TR_1) - r(S|TR_2)} \right) - 1.$$

Under A1 and A2, we easily obtain the following estimates:

$$\begin{aligned} |E_{\mathcal{M}} X_S(q) - 1| &\leq \frac{2\varepsilon(q)}{\beta} \leq \frac{2\gamma}{\beta} \text{ for all } q \in TR_1, \\ |E_{\mathcal{M}} X_S(q) - (-1)| &\leq \frac{2\varepsilon(q)}{\beta} \leq \frac{2\gamma}{\beta} \text{ for all } q \in TR_2, \end{aligned}$$

e.g., for $q \in TR_1$,

$$\begin{aligned} E_{\mathcal{M}} \frac{1_S - r(S|TR_2)}{r(S|TR_1) - r(S|TR_2)} &= \sum_{i=1}^k E_{\mathcal{M}} \left(\frac{1_S - r(S|TR_2)}{r(S|TR_1) - r(S|TR_2)} | \mathcal{M}_i \right) P_{\mathcal{M}}(\mathcal{M}_i) \\ &= \sum_{i=1}^k E_{\mathcal{M}} \left(\frac{1_S - x_2^{(i)}}{x_1^{(i)} - x_2^{(i)}} | \mathcal{M}_i \right) P_{\mathcal{M}}(\mathcal{M}_i) \\ &= \sum_{i=1}^k \frac{P_{\mathcal{M}}(q \in S | \mathcal{M}_i) - x_2^{(i)}}{x_1^{(i)} - x_2^{(i)}} P_{\mathcal{M}}(\mathcal{M}_i), \end{aligned}$$

from which the estimate immediately follows using A2; see [2, Theorem 1], and, for $\varepsilon(q) = 0$, [11, Lemma 4].

So if γ is suitably small and β suitably large, $E_{\mathcal{M}} X_S(q)$ is close to 1 for $q \in TR_1$ and close to -1 for $q \in TR_2$, so the X_S provide weak learners in the sense of Section 2 for $(q, j) \in TR$.

4 THE CONDITIONS OF [11], [2] FOR THE GENERALIZATION TO THE TEST SETS

Having designed a classification procedure with regard to the training set, the essential question is that of its generalization properties, i.e., we ask how large the classification error on test sets is. In [11], [2], a test set of the form

$$TE = TE_1 \times \{1\} \cup TE_2 \times \{2\}$$

is considered and treated in its whole for obtaining conditions for generalization. Conditions were found in [11] such that the behavior of the weak learner on the training set is reproduced on the test set. Again these conditions were weakened in [2]. We state them in the following form, where $\delta \geq 0, \gamma^* \geq 0$.

B1. δ -Indiscernibility of TE and TR

For all $S \in \mathcal{M}, i = 1, 2$,

$$|r(S|TR_i) - r(S|TE_i)| \leq \delta.$$

In [11], $\delta = 0$ is required. As discussed in [2], $\delta = 0$ is an idealization of what happens in practice and, to obtain a small δ , a strong similarity also concerning the size of TR_i and TE_i is required.

The following assumption of [2] is as A2, but now for TE with corresponding $\mathcal{M}_1^*, \dots, \mathcal{M}_{k^*}^*, \varepsilon^*, \delta^*$, so we state it as

B2. Near Uniformity Assumption $NUA(\gamma^*, TE)$.

The formulation is as A2 with γ^* replacing γ , TE replacing TR .

In [11], $\gamma^* = 0$ is required.

Under A1 and B1, TE clearly fulfills A1 with $\beta^* = \delta - 2\beta$, where we assume $\delta > 2\beta$. Hence for

$$X_S^* = 2 \left(\frac{1_S - r(S|TE_2)}{r(S|TE_1) - r(S|TE_2)} \right) - 1,$$

it follows, as in Section 3, that

$$\begin{aligned} |E_{\mathcal{M}} X_S^*(q) - 1| &\leq \frac{2\gamma^*}{\beta - 2\delta} \text{ for all } q \in TE_1, \\ |E_{\mathcal{M}} X_S^*(q) - (-1)| &\leq \frac{2\gamma^*}{\beta - 2\delta} \text{ for all } q \in TE_2. \end{aligned}$$

A1 and A2 immediately imply

$$|X_S(q) - X_S^*(q)| \leq \frac{4\delta}{\beta(\beta - 2\delta)}$$

for all $q \in F$; thus

$$|E_{\mathcal{M}} X_S(q) - 1| \leq \frac{2\gamma^*}{\beta - 2\delta} + \frac{4\delta}{\beta(\beta - 2\delta)}$$

for all $q \in TE_1$,

$$|E_{\mathcal{M}} X_S(q) - (-1)| \leq \frac{2\gamma^*}{\beta - 2\delta} + \frac{4\delta}{\beta(\beta - 2\delta)}$$

for all $q \in TE_2$.

So for suitably small γ^*, δ , in particular for $\gamma^* = \delta = 0$, the X_S are weak learners for all $(q, j) \in TE$; hence the generalization property holds.

5 A NEW CONDITION FOR THE GENERALIZATION PROPERTY

To apply SD, we start with a training set TR . With regard to this training set, a class $\mathcal{M}$ is constructed in such a manner that conditions A1 and A2 are fulfilled for suitable β and ε . A concrete algorithm for such a construction is given in [2]. In order that $\mathcal{M}$ provides weak learners on a test set TE , conditions B1 and B2 are used as explained in Section 4. These conditions seem to be hard to check for a TE and require a strong similarity of TR and TE .

In applications of pattern recognition, we usually have a large set TR of suitably generated training data, but may have an online classification problem. This means that we classify an incoming data stream where there is a changing TE which at any instant may consist of a few or even of one element of F and then B1 and B2 are no longer meaningful.

So, it seems of interest to find a condition for the generalization property of SD which only concerns the relation of a single sample to be classified with regard to the training set TR . We now provide such a condition. We let $\alpha \geq 0$ and consider a single $(q, i) \in F \times \{1, 2\}$.

C. Generalization property $GP((q, i), TR, \alpha)$

$$\begin{aligned} \min_{q \in TR_i} [P_{\mathcal{M}}(q \in S, \bar{q} \notin S) + P_{\mathcal{M}}(q \notin S, \bar{q} \in S)] \\ = \alpha(q, i) \leq \alpha. \end{aligned}$$

When we have an observation $q \in F$ belonging to class i , we expect it to be close, in some topological sense, to TR_i . So, it should be hard for sets S with good generalization properties to distinguish q from those elements of TR_i which are close to q . This is formalized in C. Note that, for $TR'_i \supseteq TR_i$ with corresponding $\alpha'(q, i)$, the minimum is taken over more elements so that we have

$$\begin{aligned} \alpha'(q, i) &= \min_{\bar{q} \in TR'_i} [P_{\mathcal{M}}(q \in S, \bar{q} \notin S) + P_{\mathcal{M}}(q \notin S, \bar{q} \in S)] \\ &\leq \min_{\bar{q} \in TR_i} [P_{\mathcal{M}}(q \in S, \bar{q} \notin S) + P_{\mathcal{M}}(q \notin S, \bar{q} \in S)] \\ &= \alpha(q, i). \end{aligned}$$

Hence, when the training set gets larger, the constants $\alpha(q, i)$ become smaller.

Our following result shows that the weak learning properties of X_S with regard to (q, i) become better the smaller $\alpha(q, i)$ is. This means that the generalization properties of X_S become better with increasing training sequence, of course, controlling β, γ from A1, A2, and point out that overtraining with regard to increasing training sequences should not occur. As we are averaging over the number n of resamples, an increase in n will not cause overtraining. Condition C might provide a mathematical explanation for the overtraining-resistance which has been found in practice.

Theorem 5.1. Assume A1, A2 for TR . Let $(q, i) \in F \times \{1, 2\}$ and assume C. Then with β, γ from A1, A2, $\alpha(q, i), \alpha$ from C:

$$\begin{aligned} |E_{\mathcal{M}} X_S(q) - 1| &\leq 2 \frac{\gamma + \alpha(q, i)}{\beta} \leq 2 \frac{\gamma + \alpha}{\beta} \text{ for } i = 1, \\ |E_{\mathcal{M}} X_S(q) - (-1)| &\leq 2 \frac{\gamma + \alpha(q, i)}{\beta} \leq 2 \frac{\gamma + \alpha}{\beta} \text{ for } i = 2. \end{aligned}$$

Proof. Let $i = 1$. Choose $\bar{q} \in TR_1$ such that

$$\alpha(q, i) = P_{\mathcal{M}}(q \in S, \bar{q} \notin S) + P_{\mathcal{M}}(q \notin S, \bar{q} \in S).$$

Then,

$$\begin{aligned} E_{\mathcal{M}} X_S(q) &= 2 \left[E_{\mathcal{M}} \left(\frac{1_S(q) - r(S|TR_2)}{r(S|TR_1) - r(S|TR_2)} \right) \right] - 1 \\ &= 2 \left[E_{\mathcal{M}} \left(\frac{1_S(\bar{q}) - r(S|TR_2)}{r(S|TR_1) - r(S|TR_2)} \right) \right] - 1 \\ &\quad + E_{\mathcal{M}} \left(\frac{1_{\{q \in S, \bar{q} \notin S\}} - 1_{\{q \notin S, \bar{q} \in S\}}}{r(S|TR_1) - r(S|TR_2)} \right) - 1 \\ &= E_{\mathcal{M}} X_S(\bar{q}) \\ &\quad + 2 E_{\mathcal{M}} \left(\frac{1_{\{q \in S, \bar{q} \notin S\}} - 1_{\{q \notin S, \bar{q} \in S\}}}{r(S|TR_1) - r(S|TR_2)} \right). \end{aligned}$$

Hence, with A1, A2, and C:

$$\begin{aligned} |E_{\mathcal{M}} X_S(q) - 1| &\leq |E_{\mathcal{M}} X_S(\bar{q}) - 1| \\ &\quad + 2 \left| E_{\mathcal{M}} \left(\frac{1_{\{q \in S, \bar{q} \notin S\}} - 1_{\{q \notin S, \bar{q} \in S\}}}{r(S|TR_1) - r(S|TR_2)} \right) \right| \\ &\leq 2 \frac{\varepsilon(\bar{q})}{\beta} + 2 E_{\mathcal{M}} \left| \frac{1_{\{q \in S, \bar{q} \notin S\}} - 1_{\{q \notin S, \bar{q} \in S\}}}{r(S|TR_1) - r(S|TR_2)} \right| \\ &\leq 2 \frac{\varepsilon(\bar{q})}{\beta} + \frac{2}{\beta} E_{\mathcal{M}} |1_{\{q \in S, \bar{q} \notin S\}} - 1_{\{q \notin S, \bar{q} \in S\}}| \\ &= 2 \frac{\varepsilon(\bar{q})}{\beta} + \frac{2}{\beta} [P_{\mathcal{M}}(q \in S, \bar{q} \notin S) \\ &\quad + P_{\mathcal{M}}(q \notin S, \bar{q} \in S)] \\ &= 2 \frac{\varepsilon(\bar{q})}{\beta} + 2 \frac{\alpha(q, i)}{\beta} \leq 2 \frac{\gamma + \alpha}{\beta}. \end{aligned}$$

The case $i = 2$ is similar. $\square$

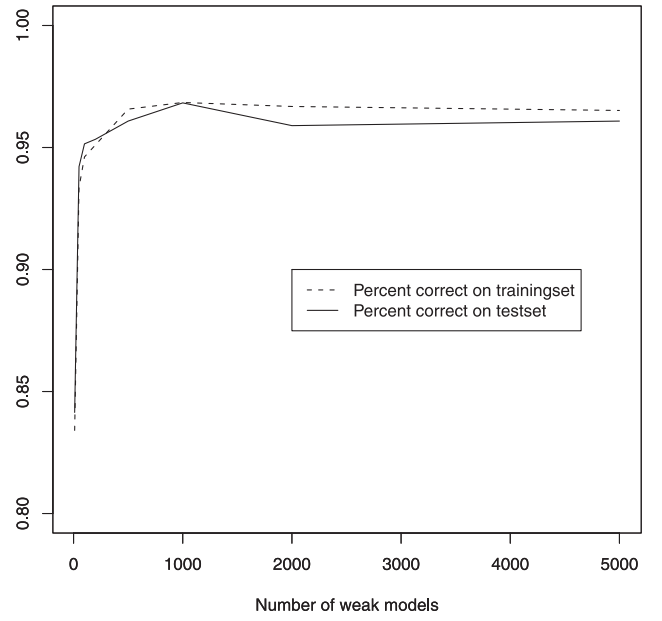


Fig. 1. SD performance.

Remarks.

1. Note that β, γ are constants which relate only to the training set, and, by a good construction of $\mathcal{M}$, see e.g., [12], [2], may be kept suitably controlled. Hence, the generalization properties depend on the constants $\alpha(q, i)$. Now if q is an observation of class i , which is not too untypical for its class, we may expect to have a $\bar{q}$ from TR_i close to it. This, for a suitable class $\mathcal{M}$, is already enough to have a small $\alpha(q, i)$. Furthermore, the availability of a close $\bar{q}$ becomes better the larger TR_i gets.
2. In [11], [2], the variance of $X_S(q)$ is also bounded, using A1, A2, and A1, A2, B1, B2, respectively. This may be done as well in our setting from A1, A2, and C. The details are omitted, see [9].

6 AN EXPERIMENTAL ILLUSTRATION

In this section, we report an experimental result in order to illustrate the method of SD. For more in-depth experimental studies, we refer to [12], [2]. We used the US Postal Service data set of handwritten digits, available at [13], where these digits are represented by 16×16 gray-scale matrices, so the feature space is 256-dimensional euclidean space. In the cited references, the elements of $\mathcal{M}$ are rectangles and unions of rectangles. To illustrate that SD is not restricted to these geometrical shapes, we used euclidean balls S for our experiments with points of the training sets as midpoints and randomly chosen diameters. We looked at two class problems to distinguish between digits i and j , the training set consisting of all gray-value matrices for these two digits. To generate a ball, we fixed parameters β, c, d, ℓ . A midpoint in $TR = TR_i \cup TR_j$ was drawn at random, then a radius $r \leq \ell$ was chosen, both according to a uniform distribution. This random ball S was accepted to form a weak learner if $|r(S|TR_i) - r(S|TR_j)| \geq \beta$ and $c \leq |r(S|TR_i \cup TR_j)| \leq d$. The first condition is related to A1, the second condition removes balls which are not suitable for TR . Note that this is computationally very simple and fast and no further steps were involved as it was not our intention to build a sophisticated classifier but to illustrate how SD works. For the following experimental result, we used the digits $i = 0$ with 1,194 examples in the training set, 359 examples in the test set, $j = 9$ with 644 examples in the training set, 177 examples in the

TABLE 1

Number of weak models	Percent correct on trainingset	Percent correct on testset
10	83.41	84.14
50	93.36	94.22
100	94.61	95.15
500	96.57	96.08
1000	96.84	96.82
5000	96.52	96.08

test set, and $\beta = 0.2, c = 0.01, d = 0.99, \ell = 30$. Fig. 1 and Table 1 illustrate the exponential decrease of misclassification and the resistance against overtraining when the number of weak models, i.e., resamples, is increased.

7 BAGGING, BOOSTING, AND SD

For a comparison, we look at two class problems and describe classifiers by classification functions h as before, a classifier being a classification function which only takes the values -1 and $+1$. All three methods start with some space $\mathcal{H}$ of classification functions or classifiers, often called weak learners in the boosting context, and look for some linear combination $\sum_i \alpha_i h_i$ in the linear span of $\mathcal{H}$ with improved classification properties. The set $\mathcal{H}$ may consist of certain classification trees for bagging and boosting, but is of a very different nature in SD, as described in Section 3 of this paper.

7.1 Bagging

Bagging, as introduced in [1], stems from the statistical notion of bootstrapping, which stands for resampling of the data to form bootstrap samples. Here, the training set TR is resampled with replacement to form a collection of B bootstrap training sets $TR^1, \dots, TR^B$. To each TR^i , a classification function h_i , $i = 1, \dots, B$, is fitted, and the bagging classification function is given by $\frac{1}{B} \sum_i h_i$. So if each h_i is a classifier, we have a majority vote. Bagging may be seen as a way of reducing the variance in unstable procedures like classification trees, but often will not lead from weak learners to strong learners.

7.2 Boosting

Boosting goes back to [4], [5]. The original AdaBoost sequentially finds weak classifiers h_m and coefficients α_m , $m = 1, \dots, M$, and arrives at the classification function $\sum_m \alpha_m h_m$, hence a weighted majority vote. At each boosting step m , weights $w_k^{(m)}$ are applied to each of the $n = 1, \dots, N$ training observations. Then, a classifier h_m is fitted to the training set, minimizing the empirical error weighted by the $w_k^{(m)}$. This classifier is supplied with a coefficient α_m depending on the weighted empirical error, which gives more importance in the majority vote to accurate classifiers h_m . Then, the weights $w_k^{(m)}$ are updated to form $w_k^{(m+1)}$, where misclassified observations in step m have their weights increased, correctly classified observations have their weights decreased, and the $m+1$ th round is started. In [6], a new view of boosting was obtained and has led to a wealth of new developments. It was shown that boosting could be seen as a sequential minimization procedure in the linear span of $\mathcal{H}$. At each step m , having already computed f_{m-1} , we minimize the empirical loss for $f_{m-1} + \alpha h$ over $h \in \mathcal{H}$, $\alpha \in \mathbb{R}$, where misclassification is now measured by a certain loss function. The resulting minimizer $\alpha_m h_m$ is added to form $f_m = f_{m-1} + \alpha_m h_m$. For exponential loss function, this results in AdaBoost. Boosting has been very successful for obtaining strong learners from weak learners, but some care has to be taken when choosing M , as overfitting has been detected for large M .

7.3 Stochastic Discrimination

As bagging, SD is a resampling procedure, but of a completely different nature. There is no resampling of the data; instead, at each step $i = 1, \dots, n$, we resample the classification functions from the space of weak learners, so, for each i , a classification function h_i is obtained randomly from $\mathcal{H}$ with respect to the resampling distribution $P_{\mathcal{H}}$. The resulting classification function then is $\frac{1}{n} \sum_i h_i$ and provides a strong learner from the weak learners h_i .

Whereas in bagging and boosting, at each step a fitting procedure, e.g., empirical loss minimization, is involved, this is not the case for SD, where only resampling, with distribution fixed in advance, is required. The fitting of SD to the training set is incorporated in the design of the sets $\mathcal{H}$, respectively, $\mathcal{M}$, and, as described in [12], [2], general principles for this design are available which work for a large class of problems. As a conclusion of this discussion we may remark that, although the three methods are ensemble methods, SD differs very markedly from bagging and boosting.

ACKNOWLEDGMENTS

The authors thank the reviewers for their valuable comments and suggestions.

REFERENCES

- [1] L. Breiman, "Bagging Predictors," *Machine Learning*, vol. 26, pp. 123-140, 1996.
- [2] D. Chen, P. Huang, and X. Cheng, "A Concrete Statistical Realization of Kleinberg's Stochastic Discrimination for Pattern Recognition. I. Two-Class Classification," *Annals of Statistics*, vol. 31, pp. 1393-1412, 2003.
- [3] R. Durrett, *Probability: Theory and Examples*. Wadsworth, 1991.
- [4] Y. Freund, "Boosting a Weak Learning Algorithm by Majority," *Information and Computation*, vol. 121, pp. 256-285, 1995.
- [5] Y. Freund and R. Schapire, "Experiments with a New Boosting Algorithm," *Proc. 13th Int'l Conf. Machine Learning*, pp. 148-156, 1996.
- [6] J. Friedman, T. Hastie, and R. Tibshirani, "Additive Logistic Regression: A Statistical View of Boosting (with Discussion)," *Annals of Statistics*, vol. 28, pp. 337-407, 2000.
- [7] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.
- [8] T.K. Ho, "The Random Subspace Method for Constructing Decision Forests," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832-844, Aug. 1998.
- [9] A. Irle and J. Kauschke, "On Kleinberg's Stochastic Discrimination Procedure," technical report, Univ. of Kiel, 2009.
- [10] E.M. Kleinberg, "Stochastic Discrimination," *Annals of Math. and Artificial Intelligence*, vol. 1, pp. 207-239, 1990.
- [11] E.M. Kleinberg, "An Overtraining-Resistant Stochastic Modeling Method for Pattern Recognition," *Annals of Statistics*, vol. 24, pp. 2319-2349, 1996.
- [12] E.M. Kleinberg, "On the Algorithmic Implementation of Stochastic Discrimination," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 5, pp. 473-490, May 2000.
- [13] <http://www-stat.stanford.edu/tibs/ElemStatLearn/data.html>, 2010.

► For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.