Machine Learning, 45, 5–32, 2001

Random Forests

LEO BREIMAN
Statistics Department, University of California, Berkeley, CA 94720

Editor: Robert E. Schapire

Abstract. Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges a.s. to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Using a random selection of features to split each node yields error rates that compare favorably to Adaboost (Y. Freund & R. Schapire, *Machine Learning: Proceedings of the Thirteenth International conference*, ∗ ∗ ∗, 148–156), but are more robust with respect to noise. Internal estimates monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. Internal estimates are also used to measure variable importance. These ideas are also applicable to regression.

Keywords: classification, regression, ensemble

1. Random forests

1.1. Introduction

Significant improvements in classification accuracy have resulted from growing an ensemble of trees and letting them vote for the most popular class. In order to grow these ensembles, often random vectors are generated that govern the growth of each tree in the ensemble. An early example is bagging (Breiman, 1996), where to grow each tree a random selection (with replacement) is made from the examples in the training set.

Another example is random split selection (Dietterich, 1998) where at each node the split is selected at random from among the K best splits. Breiman (1999) generates new training sets by randomizing the outputs in the original training set. Another approach is to select the training set from a random set of weights on the examples in the training set. Ho (1998) has written on the random subspace method which does a random selection of a subset of features to use to grow each tree.

In an important paper on writing recognition, Amit and Geman (1997) define a large number of geometric features and search over a random selection of these for the best split at each node. This latter paper has been influential in my thinking.

The common element in all of these procedures is that for the kth tree, a random vector Θ_k is generated, independent of the past random vectors $\Theta_1, \ldots, \Theta_{k-1}$ but with the same distribution; and a tree is grown using the training set and Θ_k, resulting in a classifier $h(\mathbf{x}, \Theta_k)$ where $\mathbf{x}$ is an input vector. For instance, in bagging the random vector Θ is

generated a the co m in N bo e re h ing from N dan hrough an random a he bo e , here N i n mber of e ample in he raining e . In random pli elec ion Θ con i of a n mber of independen random im eger be een 1 and K. The na re and dimen ionali of Θ depend on i e in ree con r c ion.

Af er a large n mber of ree i genera ed, he o e for he mo pop lar cla . We call he e proced re **random forests**.

Definition 1.1. A random fore i a cla i er con i ing of a collec ion of ree- r c red cla i er $\{h(\mathbf{x}, \Theta_k),\ k = 1, \ldots\}$ here he $\{\Theta_k\}$ are independen iden ically di rib ed random ec or and each ree ca a ni o e for he mo pop lar cla a inp **x**.

1.2. Outline of paper

Sec ion 2 gi e ome heore ical backgro nd for random fore . U e of he S rong La of Large N mber ho (ha he al a con erge o ha o er ing i no a problem. We gi e a impli ed and e ended er ion of he Ami and Geman (1997) anal i o ho (ha he acc rac of a random fore depend on he reng h of he indi i- d al ree cla i er and a mea re of he dependence be een hem (ee Sec ion 2 for de ni ion).

Sec ion 3 in rod ce fore ing he random elec ion of fea re a each node o de ermine he pli . An impor an q e ion i ho (man fea re o elec a each node. For g idance, in ernal e ima e of he generali a ion error, cla i er reng h and dependence are comp ed. The e are called o -of-bag e ima e and are re ie ed in Sec ion 4. Sec ion 5 and 6 gi e empirical re h for o differen form of random fea re . The r e random elec ion from he original inp ; he econd e random linear combina ion of inp . The re h compare fa orabl o Adaboo .

The re h rn o o be in en i i e o he n mber of fea re elec ed o pli each node. U all , elec ing one or o fea re gi e near op im m re h . To e plore hi and rela e i o reng h and correla ion, an empirical d i carried o in Sec ion 7.

Adaboo ha no random elemen and gro an en emble of ree b cce i e re eigh- ing of he raining e here he c rren eigh depend on he pa hi or of he en emble forma ion. B j a a de ermini ic random n mber genera or can gi e a good imi a ion of randomne , m belief i ha in i la er age Adaboo i em la ing a random fore . E idence for hi conjec re i gi en in Sec ion 8.

Impor an recen problem , i.e., medical diagno i and doc men re rie al, of en ha e he proper ha here are man inp ariable , of en in he h ndred or ho and , h each one con aining onl a mall amo m of informa ion. A ingle ree cla i er ill hen ha e acc rac onl ligh l be er han a random choice of cla . B combining ree gro n ing random fea re can prod ce impro ed acc rac . In Sec ion 9 e e perimen on a im la ed da a e h 1,000 inp ariable , 1,000 e ample in he raining e and a 4,000 e ample e e . Acc rac comparable o he Ba e ra e i achie ed.

In man applica ion , nder anding of he mechani m of he random fore black bo i needed. Sec ion 10 make a an on hi b comp ing in ernal e ima e of ariable impor ance and binding he e oge her b re e r n .

Sec ion 11 look a random fore for regre ion. A bo nd for he mean q ared gener-ali a ion error i deri ed ha ho w ha he decrea e in error from he indi id al ree in he fore depend on he correla ion be ween re id al and he mean q ared error of he indi id al ree . Empirical re l for regre ion are in Sec ion 12. Concl ding remark are gi en in Sec ion 13.

2. Characterizing the accuracy of random forests

2.1. Random forests converge

Gi en an en emble of cla i er $h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_K(\mathbf{x})$, and w ih he raining e dra n a random from he di rib ion of he random ec or Y, $\mathbf{X}$, de ne he margin f nc ion a

$$mg(\mathbf{X}, Y) = av_k I(h_k(\mathbf{X}) = Y) - \max_{j \neq Y} av_k I(h_k(\mathbf{X}) = j).$$

where $I(\cdot)$ i he indica or f nc ion. The margin mea re he e em o which he a erage n mber of o e a $\mathbf{X}, Y$ for he righ cla e ceed he a erage o e for an o her cla . The larger he margin, he more con dence in he cla i ca ion. The generali a ion error i gi en b

$$PE^* = P_{\mathbf{X},Y}(mg(\mathbf{X}, Y) < 0)$$

where he b crip $\mathbf{X}, Y$ indica e ha he probabili i o er he $\mathbf{X}, Y$ pace.

In random fore , $h_k(\mathbf{X}) = h(\mathbf{X}, \Theta_k)$. For a large n mber of ree , i follo from he S rong La of Large N mber and he ree r c re ha :

Theorem 1.2. *As the number of trees increases, for almost surely all sequences* $\Theta_{1,\ldots}$ *PE* converges to*

$$P_{\mathbf{X},Y}(P_\Theta(h(\mathbf{X}, \Theta) = Y) - \max_{j \neq Y} P_\Theta(h(\mathbf{X}, \Theta) = j) < 0). \quad (1)$$

Proof: ee Appendi I. □

Thi re l e plain w h random fore do no o er a more ree are added, b prod ce a limi ing al e of he generali a ion error.

2.2. Strength and correlation

For random fore , an pper bo nd can be deri ed for he generali a ion error in erm of w o parame er ha are mea re of ho w acc ra e he indi id al cla i er are and of he dependence be ween hem. The im erpla be ween he e w o gi e he fo nda ion for nder anding he working of random fore . We b ild on he anal i in Ami and Geman (1997).

Definition 2.1. The margin f nc ion for a random fore i

$$mr(\mathbf{X}, Y) = P_\Theta(h(\mathbf{X}, \Theta) = Y) - \max_{j \neq Y} P_\Theta(h(\mathbf{X}, \Theta) = j) \tag{2}$$

and he reng h of he e of cla i er $\{h(\mathbf{x}, \Theta)\}$ i

$$s = E_{\mathbf{X},Y} mr(\mathbf{X}, Y). \tag{3}$$

A ming $s \geq 0$, Cheb che ' ineq ali gi e

$$PE^* \leq \text{ar}(mr)/s^2 \tag{4}$$

A more re ealing e pre ion for he ariance of mr i deri ed in he follo ing: Le

$$\hat{j}(\mathbf{X}, Y) = \arg\max_{j \neq Y} P_\Theta(h(\mathbf{X}, \Theta) = j)$$

o

$$\begin{aligned} mr(\mathbf{X}, Y) &= P_\Theta(h(\mathbf{X}, \Theta) = Y) - P_\Theta(h(\mathbf{X}, \Theta) = \hat{j}(\mathbf{X}, Y)) \\ &= E_\Theta[I(h(\mathbf{X}, \Theta) = Y) - I(h(\mathbf{X}, \Theta) = \hat{j}(\mathbf{X}, Y)]. \end{aligned}$$

Definition 2.2. The ra margin f nc ion i

$$rmg(\Theta, \mathbf{X}, Y) = I(h(\mathbf{X}, \Theta) = Y) - I(h(\mathbf{X}, \Theta) = \hat{j}(\mathbf{X}, Y)).$$

Th , $mr(\mathbf{X}, Y)$ i he e pec ation of $rmg(\Theta, \mathbf{X}, Y)$ i h re pec o Θ. For an f nc ion f he iden i

$$[E_\Theta f(\Theta)]^2 = E_{\Theta,\Theta'} f(\Theta) f(\Theta')$$

hold here Θ, Θ' are independen i h he ame di rib ion, impl ing ha

$$mr(\mathbf{X}, Y)^2 = E_{\Theta,\Theta'} rmg(\Theta, \mathbf{X}, Y) rmg(\Theta', \mathbf{X}, Y) \tag{5}$$

U ing (5) gi e

$$\begin{aligned} \text{ar}(mr) &= E_{\Theta,\Theta'}(\text{co}_{\mathbf{X},Y} rmg(\Theta, \mathbf{X}, Y) rmg(\Theta', \mathbf{X}, Y)) \\ &= E_{\Theta,\Theta'}(\rho(\Theta, \Theta') sd(\Theta) sd(\Theta')) \end{aligned} \tag{6}$$

here $\rho(\Theta, \Theta')$ i he correla ion be een $rmg(\Theta, \mathbf{X}, Y)$ and $rmg(\Theta', \mathbf{X}, Y)$ holding Θ, Θ' ed and $sd(\Theta)$ i he andard de ia ion of $rmg(\Theta, \mathbf{X}, Y)$ holding Θ ed. Then,

$$\begin{aligned} \text{ar}(mr) &= \bar{\rho}(E_\Theta sd(\Theta))^2 \\ &\leq \bar{\rho} E_\Theta \text{ ar}(\Theta) \end{aligned} \tag{7}$$

where $\bar{\rho}$ is the mean value of the correlation; that is,

$$\bar{\rho} = E_{\Theta,\Theta'}(\rho(\Theta,\Theta')sd(\Theta)sd(\Theta'))/E_{\Theta,\Theta'}(sd(\Theta)sd(\Theta'))$$

Write

$$\begin{aligned} E_\Theta \text{ var}(\Theta) &\le E_\Theta(E_{\mathbf{X},Y} rmg(\Theta,\mathbf{X},Y))^2 - s^2 \\ &\le 1 - s^2. \end{aligned} \tag{8}$$

Putting (4), (7), and (8) together yields:

Theorem 2.3. *An upper bound for the generalization error is given by*

$$PE^* \le \bar{\rho}(1-s^2)/s^2.$$

Although the bound is likely to be loose, it fits the same general inverse form for random forests as a VC-type bound does for other types of classifiers. It shows that the two ingredients involved in the generalization error for random forests are the strength of the individual classifiers in the forest, and the correlation between them in terms of the raw margin function. The c/2 ratio is the correlation divided by the square of the strength. Understanding the functioning of random forests, this ratio will be a helpful guide: the smaller it is, the better.

Definition 2.4. The c/2 ratio for a random forest is defined as

$$c/s2 = \bar{\rho}/s^2.$$

There are implications in the two classification. The margin function is

$$mr(\mathbf{X},Y) = 2P_\Theta(h(\mathbf{X},\Theta) = Y) - 1$$

The requirement that the strength is positive (see (4)) becomes similar to the familiar weak learning condition $E_{\mathbf{X},Y}P_\Theta(h(\mathbf{X},\Theta) = Y) > .5$. The raw margin function is $2I(h(\mathbf{X},\Theta) = Y) - 1$ and the correlation $\bar{\rho}$ is between $I(h(\mathbf{X},\Theta) = Y)$ and $I(h(\mathbf{X},\Theta') = Y)$. In particular, if the values for Y are taken to be $+1$ and -1, then

$$\bar{\rho} = E_{\Theta,\Theta'}[\rho(h(\cdot,\Theta), h(\cdot,\Theta'))]$$

so that $\bar{\rho}$ is the correlation between two different members of the forest averaged over the Θ,Θ' distribution.

For more than two classes, the measure of strength defined in (3) depends on the forest as well as the individual tree inference in it the forest that determine $\hat{j}(\mathbf{X},Y)$. Another approach

i po ible. Wri e

$$PE^* = P_{\mathbf{X},Y}(P_\Theta(h(\mathbf{X},\Theta) = Y) - \max_{j \neq Y} P_\Theta(h(\mathbf{X},\Theta) = j) < 0)$$
$$\leq \sum_j P_{\mathbf{X},Y}(P_\Theta(h(\mathbf{X},\Theta) = Y) - P_\Theta(h(\mathbf{X},\Theta) = j) < 0).$$

De ne

$$s_j = E_{\mathbf{X},Y}(P_\Theta(h(\mathbf{X},\Theta) = Y) - P_\Theta(h(\mathbf{X},\Theta) = j))$$

o be he reng h of he e of cla i er $\{h(\mathbf{x},\Theta)\}$ rela i e o cla j. No e ha hi de ni ion of reng h doe no depend on he fore . U ing Cheb he ' ineq ali , a ming all $s_{j>0}$ lead o

$$PE^* \leq \sum_j \text{ ar}(P_\Theta(h(\mathbf{X},\Theta) = Y) - P_\Theta(h(\mathbf{X},\Theta) = j))s_j^2 \tag{9}$$

and ing iden ie imilar o ho e ed in deri ing (7), he ariance in (9) can be e pre ed in erm of a erage correla ion . I did no e e ima e of he q an i ie in (9) in o r empirical d b hink he o ld be in ere ing in a m ltiple cla problem.

3. Using random features

Some random fore repor ed in he li era re ha e con i en l lo er generali a ion error han o her . For inance, random pli elec ion (Die errich, 1998) doe be er han bagging. Breiman' im rod c ion of random noi e in o he o p (Breiman, 1998c) al o doe be er. B none of he e he e hree fore do a ell a Adaboo (Fre nd & Schapire, 1996) or o her algori hm ha ork b adap i e re eigh ing (arcing) of he raining e (ee Breiman, 1998b; Die errich, 1998; Ba er & Koha i, 1999).

To impro e acc rac , he randomne inje ed ha o minimi e he correla ion $\bar{\rho}$ hile main aining reng h. The fore ied here con i of ing randoml elec ed inp or combina ion of inp a each node o gro each ree. The re l ing fore gi e acc rac ha compare fa orabl i h Adaboo . Thi cla of proced re ha de irable charac eri ic :

i I acc rac i a good a Adaboo and ome ime be er.
ii I ' rela i el rob o o lier and noi e.
iii I ' fa er han bagging or boo ing.
i I gi e ef l in ernal e ima e of error, reng h, correla ion and ariable impor ance.
I ' imple and ea il paralleli ed.

Ami and Geman (1997) gro hallo ree for hand ri en charac er recogni ion ing random elec ion from a large n mber of geome ricall de ned fea re o de ne he pli a each node. Al ho gh m implemen a ion i differen and no problem peci c, i a heir ork ha pro ided he an for m idea .

3.1. Using out-of-bag estimates to monitor error, strength, and correlation

In m e perimen u ing h random fore , bagging i ed in andem u ing h random fea re elec ion. Each ne raining e i dra n, u ing h replacemen , from he original raining e . Then a ree i gro n on he ne raining e ing random fea re elec ion. The ree gro n are no pr ned.

There are o rea on for ing bagging. The r i ha he e of bagging eem o enhance acc rac hen random fea re are ed. The econd i ha bagging can be ed o gi e ongoing e ima e of he generali a ion error (PE*) of he combined en emble of ree , a ell a e ima e for he reng h and correla ion. The e e ima e are done o -of-bag, hich i e plained a follo .

A me a me hod for con r c ing a cla i er from an raining e . Gi en a pec i c raining e T, form boo rap raining e T_k, con r c cla i er $h(\mathbf{x}, T_k)$ and le he e o e o form he bagged predic or. For each $y, \mathbf{x}$ in he raining e , aggrega e he o e onl o er ho e cla i er for hich T_k doe no con aining $y, \mathbf{x}$. Call hi he o -of-bag cla i er. Then he o -of-bag e ima e for he generali a ion error i he error ra e of he o -of-bag cla i er on he raining e .

Tib hirani (1996) and Wolpen and Macread (1996), propo ed ing o -of-bag e i- ma e a an ingredien in e ima e of generali a ion error. Wolpen and Macread ork on regre ion pe problem and propo ed a n mber of me hod for e ima ing he gen- erali a ion error of bagged predic or . Tib hirani ed o -of-bag e ima e of ariance o e ima e generali a ion error for arbi rar cla i er . The d of error e ima e for bagged cla i er in Breiman (1996b), gi e empirical e idence o ho ha he o - of-bag e ima e i a acc ra e a ing a e e of he ame i e a he raining e . Therefore, ing he o -of-bag error e ima e remo e he need for a e a ide e e .

In each boo rap raining e , abo one- hird of he in ance are lef o . Therefore, he o -of-bag e ima e are ba ed on combining onl abo one- hird a man cla i er a in he ongoing main combina ion. Since he error ra e decrea e a he n mber of combina ion increa e , he o -of-bag e ima e ill end o o ere ima e he c rren error ra e. To ge nbia ed o -of-bag e ima e , i i nece ar o r n pa he poin here he e e error con erge . B nlike cro - alida ion, here bia i pre en b i e en nkno n, he o -of-bag e ima e are nbia ed.

S reng h and correla ion can al o be e ima ed ing o -of-bag me hod . Thi gi e in ernal e ima e ha are helpf l in nder anding cla i ca ion acc rac and ho o impro e i . The de ail are gi en in Appendi II. Ano her applica ion i o he mea re of ariable impor ance (ee Sec ion 10).

4. Random forests using random input selection

The imple random fore ith random fea re i formed b elec ing a random, a each node, a mall gro p of inp ariable o pli on. Gro he ree ing CART me hodolog o ma im m i e and do no pr ne. Deno e hi proced re b Fore -RI. The i e F of he gro p i ed. T o al e of F ere ried. The r ed onl one randoml elec ed

Table 1. Da a e mmar .

Da a e	Train i e	Te i e	Inp	Cla e
Gla	214		9	6
Brea cancer	699		9	2
Diabe e	768		8	2
Sonar	208		60	2
Vo el	990		10	11
Iono phere	351		34	2
Vehicle	846		18	4
So bean	685		35	19
German credi	1000		24	2
Image	2310		19	7
Ecoli	336		7	8
Vo e	435		16	2
Li er	345		6	2
Le er	15000	5000	16	26
Sa -image	4435	2000	36	6
Zip-code	7291	2007	256	10
Wa eform	300	3000	21	3
T onorm	300	3000	20	2
Threenorm	300	3000	20	2
Ringnorm	300	3000	20	2

ariable, i.e., $F = 1$. The econd ook F o be he r im eger le han $\log_2 M + 1$, here M i he n mber of inp .

M e perimen e 13 maller i ed da a e from he UCI repo i or , 3 larger e epara ed im o raining and e e and 4 m he ic da a e . The r 10 e ere elec ed beca e I had ed hem in pa re earch. Table 1 gi e a brief mmar .

On each of he 13 maller i ed da a e , he follo ing proced re a ed: a random 10% of he da a a e a ide. On he remaining da a, random fore a r n ice, gro ing and combining 100 ree once h $F = 1$, and he cond ime h $F = im (\log_2 M + 1)$. The e a ide 10% a hen p do n each fore o ge a e e error for bo h. The e e error elec ed corre ponded o he lo er al e of he o -of-bag e ima e in he o r n . Thi a repea ed 100 ime and he e e error a eraged. The ame proced re a follo ed for he Adaboo r n hich are ba ed on combining 50 ree .

The e for 100 ree in random fore and 50 for Adaboo come from o o rce . The o -of-bag e ima e are ba ed on onl abo a hird a man ree a are in he fore . To ge reliable e ima e I op ed for 100 ree . The cond con idera ion i ha gro ing random fore i man ime fa er han gro ing he ree ba ed on all inp needed in Adaboo . Gro ing he 100 ree in random fore a con iderabl q icker han he 50 ree for Adaboo .

Table 2. Te ˙ e˛ error (%).

Da˛a e˛	Adaboo˙	Selec˛ion	Fore˙-RI ingle inp˙	One˙ree
Gla	22.0	20.6	21.2	36.9
Brea˙ cancer	3.2	2.9	2.7	6.3
Diabe˛e	26.6	24.2	24.3	33.1
Sonar	15.6	15.9	18.0	31.7
Vo˝el	4.1	3.4	3.3	30.4
Iono phere	6.4	7.1	7.5	12.7
Vehicle	23.2	25.8	26.4	33.1
German credi˛	23.5	24.4	26.2	33.3
Image	1.6	2.1	2.7	6.4
Ecoli	14.8	12.8	13.0	24.5
Vo˛e	4.8	4.1	4.6	7.4
Li er	30.7	25.1	24.7	40.6
Le˙er	3.4	3.5	4.7	19.8
Sa˛-image	8.8	8.6	10.5	17.2
Zip-code	6.2	6.3	7.8	20.6
Wa eform	17.8	17.2	17.3	34.0
T˝onorm	4.9	3.9	3.9	24.7
Threenorm	18.8	17.5	17.5	38.4
Ringnorm	6.9	4.9	4.9	25.7

In ˙he r n on ˙he larger da˛a e˛ , ˙he random fore˙ re l˛ for ˙he˝ r ˙ ˝o da˛a e˛ ˝ere ba ed on combining 100 ˙ree ; ˙he ip-code proced re combined 200. For Adaboo˙ , 50 ˙ree ˝ere combined for ˙he˝ r ˙ ˝hree da˛a e˛ and 100 for ip-code. The m˙he˙ic da˛a ˝a de cribed in Breiman (1996) and al o ed in Schapire e˛ al. (1997). There˝ere 50 r n . In each r n, a ne˝ ˙raining e˛ of i e 300 and ˙e ˙ e˛ of i e 3000˝ere genera˛ed. In random fore˙ 100 ˙ree ˝ere combined in each r n 50 in Adaboo˙. The re l˛ of ˙he e r n are gi en in Table 2.

The econd col mn are ˙he re l˛ elec˛ed from ˙he ˝o gro p i e b mean of lo˝e ˙ o ˙-of-bag error. The ˙hird col mn i ˙he ˙e ˛ error ing j ˙ one random fea˛ re ˛o gro˝ ˙he ˙ree . The fo rth col mn con˙ain ˙he o ˙-of-bag e ˙ima˛e of ˙he generali a˛ion error of ˙he indi id al ˙ree in ˙he fore˙ comp ˙ed for ˙he be ˙ e˛ ing (ingle or elec˛ion). Thi e ˙ima˛e i comp ˙ed b ing ˙he lef˛-o ˙ in ˙ance a a˙e ˙ e˛ in each ˙ree gro˝n and a eraging ˙he re l˛ o er all ˙ree in ˙he fore˙ .

The error ra˛e ing random inp ˙ elec˛ion compare fa orabl ˝i˙h Adaboo˙. The compari on migh˛ be e en more fa orable if ˙he earch i o er more al e of F in ˙ead of ˙he pre e˛ ˝o. B ˙ ˙he proced re i no˛ o erl en i i e ˙o ˙he al e of F. The a erage ab ol ˙e difference be˝een ˙he error ra˛e ing $F = 1$ and ˙he higher al e of F i le ˙han 1%. The difference i mo ˙ prono nced on ˙he ˙hree large da˛a e˛ .

Table 3. Te y e, error (%).

		Fore y-RC		
Da a e	Adaboo y	Selec ion	T o fea re	One ree
Gla	22.0	24.4	23.5	42.4
Brea y cancer	3.2	3.1	2.9	5.8
Diabe e	26.6	23.0	23.1	32.1
Sonar	15.6	13.6	13.8	31.7
Vo el	4.1	3.3	3.3	30.4
Iono phere	6.4	5.5	5.7	14.2
Vehicle	23.2	23.1	22.8	39.1
German credi	23.5	22.8	23.8	32.6
Image	1.6	1.6	1.8	6.0
Ecoli	14.8	12.9	12.4	25.3
Vo e	4.8	4.1	4.0	8.6
Li er	30.7	27.3	27.2	40.3
Le er	3.4	3.4	4.1	23.8
Sa -image	8.8	9.1	10.2	17.3
Zip-code	6.2	6.2	7.2	22.7
Wa eform	17.8	16.0	16.1	33.2
T onorm	4.9	3.8	3.9	20.9
Threenorm	18.8	16.8	16.9	34.8
Ringnorm	6.9	4.8	4.6	24.6

8.5%, on he le er da a o 3.0%, b he ip-code e y e, error did no decrea e. Ac ing on an informed h nch, I ried Fore y-RI in h $F = 25$. The ip-code e y e, error dropped o 5.8%. The e are he lo e y e, error o far achie ed on he e hree da a e b ree en emble .

5.1. Categorical variables

Some or all of he inp y ariable ma be ca egorical and ince e eam o de ne addi i e combina ion of ariable , e need o de ne ho ca egorical ill be rea ed o he can be combined i h n merical ariable . M approach i ha each ime a ca egorical ariable i elec ed o pli on a a node, o elec a random b e of he ca egorie of he ariable, and de ne a b in e ariable ha i one hen he ca egorical al e of he ariable i in he b e and ero o ide.

Since a ca egorical ariable i h I al e can be coded in o $I-1$ d mm $0-1$ ariable , e make he ariable $I-1$ ime a probable a a n meric ariable o be elec ed in node pli ing. When man of he ariable are ca egorical, ing a lo al e of F re h in lo correla ion, b al o lo reng h. F m be increa ed o abo o-hree ime in $(\log_2 M + 1)$ o ge eno gh reng h o pro ide good e y e, acc rac .

For in ance, on he DNA da a e ha ing 60 fo r- al ed ca egorical al e , 2,000 e am- ple in he raining e and 1,186 in he e e , ing Fore -RI h $F = 20$ ga e a e e error ra e of 3.6% (4.2% for Adaboo). The o bean da a ha 685 e ample , 35 ariable , 19 cla e , and 15 ca egorical ariable . U ing Fore -RI h $F = 12$ gi e a e e error of 5.3% (5.8% for Adaboo). U ing Fore -RC h combina ion of 3 and $F = 8$ gi e an error of 5.5%.

One ad am age of hi approach i ha i ge aro nd he dif c l of ha o do h ca egorical ha ha e man al e . In he o-cla problem, hi can be a oided b ing he de ice propo ed in Breiman e al. (1985) hich red ce he earch for he be ca egorical pli o an O(I) comp a ion. For more cla e , he earch for he be ca egorical pli i an O(2^{I-1}) comp a ion. In he random fore implemen a ion, he comp a ion for an ca egorical ariable in ol e onl he elec ion of a random b e of he ca egorie .

6. Empirical results on strength and correlation

The p rpo e of hi ec ion i o look a he effec of reng h and correla ion on he gener- ali a ion error. Ano her a pec ha e am ed o ge more nder anding of a he lack of en i i in in the generali a ion error o he gro p i e F. To cond c an empirical d of he effec of reng h and correla ion in a arie of da a e , o -of-bag e ima e of he reng h and correla ion, a de cribed in Sec ion 3.1, ere ed.

We begin b r nning Fore -RI on he onar da a (60 inp , 208 e ample) ing from 1 o 50 inp . In each i era ion, 10% of he da a a pli off a e e . Then F, he n mber of random inp elec ed a each node, a aried from 1 o 50. For each al e of F, 100 ree ere gro n o form a random fore and he erminal al e of e e error, reng h, correla ion, e c. recorded. Eigh i era ion ere done, each ime remo ing a random 10% of he da a for e a e e , and all re l a eraged o er he 80 repe i ion . Al oge her, 400,000 ree ere gro n in hi e perimen .

The op graph of g re 1, plo he al e of reng h and correla ion . F. The re l i fa cina ing. Pa abo $F = 4$ he reng h remain con an; adding more inp doe no help. B he correla ion con in e o increa e. The econd graph plo he e e error and he o -of-bag e ima e of he generali a ion error again F. The o -of-bag e ima e are more able. Bo h ho he ame beha ior a mall drop from $F = 1$ o o F abo 4 8, and hen a general, grad al increa e. Thi increa e in error allie h he beginning of he con anc region for he reng h.

Fig re 2 ha plo for imilar r n on he brea da a e here fea re con i ing of random combina ion of hree inp are ed. The n mber of he e fea re a aried from 1 o 25. Again, he correla ion ho a lo ri e, hile he reng h a in all con an , o ha he minim m error i a $F = 1$. The rpri e in he e o g re i he rela i e con anc of he reng h. Since he correla ion are lo l b ead il increa ing, he lo e error occ r hen onl a fe inp or fea re are ed.

Since he larger da a e eemed o ha e a differen beha ior han he maller, e ran a imilar e perimen on he a ellite da a e . The n mber of fea re , each con i ing of a random m of o inp , a aried from 1 o 25, and for each, 100 cla i er ere combined. The re l are ho n in g re 3. The re l differ from ho e on he maller

Figure 1. Effec of n mber of inp on onar da a.

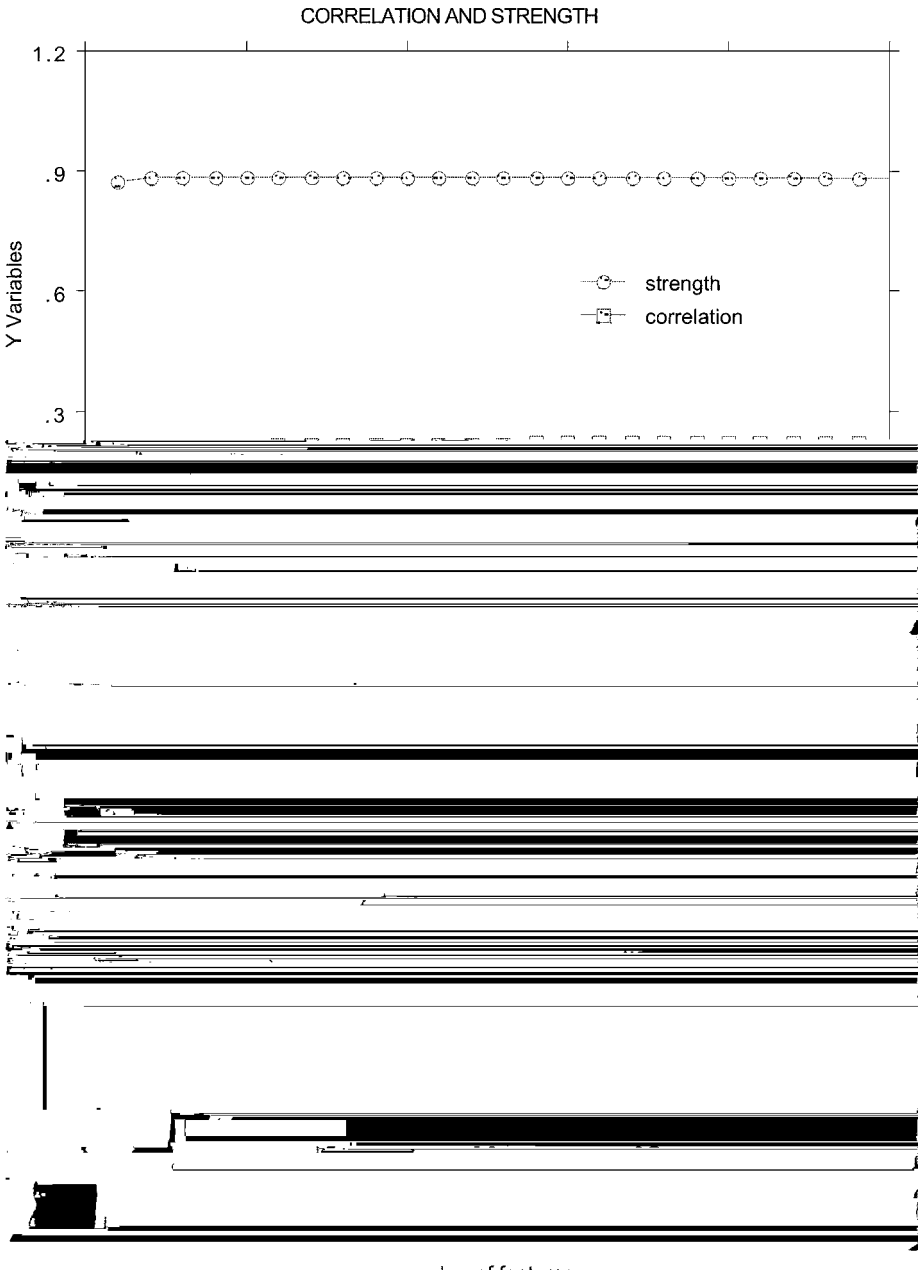

Figure 2. Effect of the number of features on the breast data set.

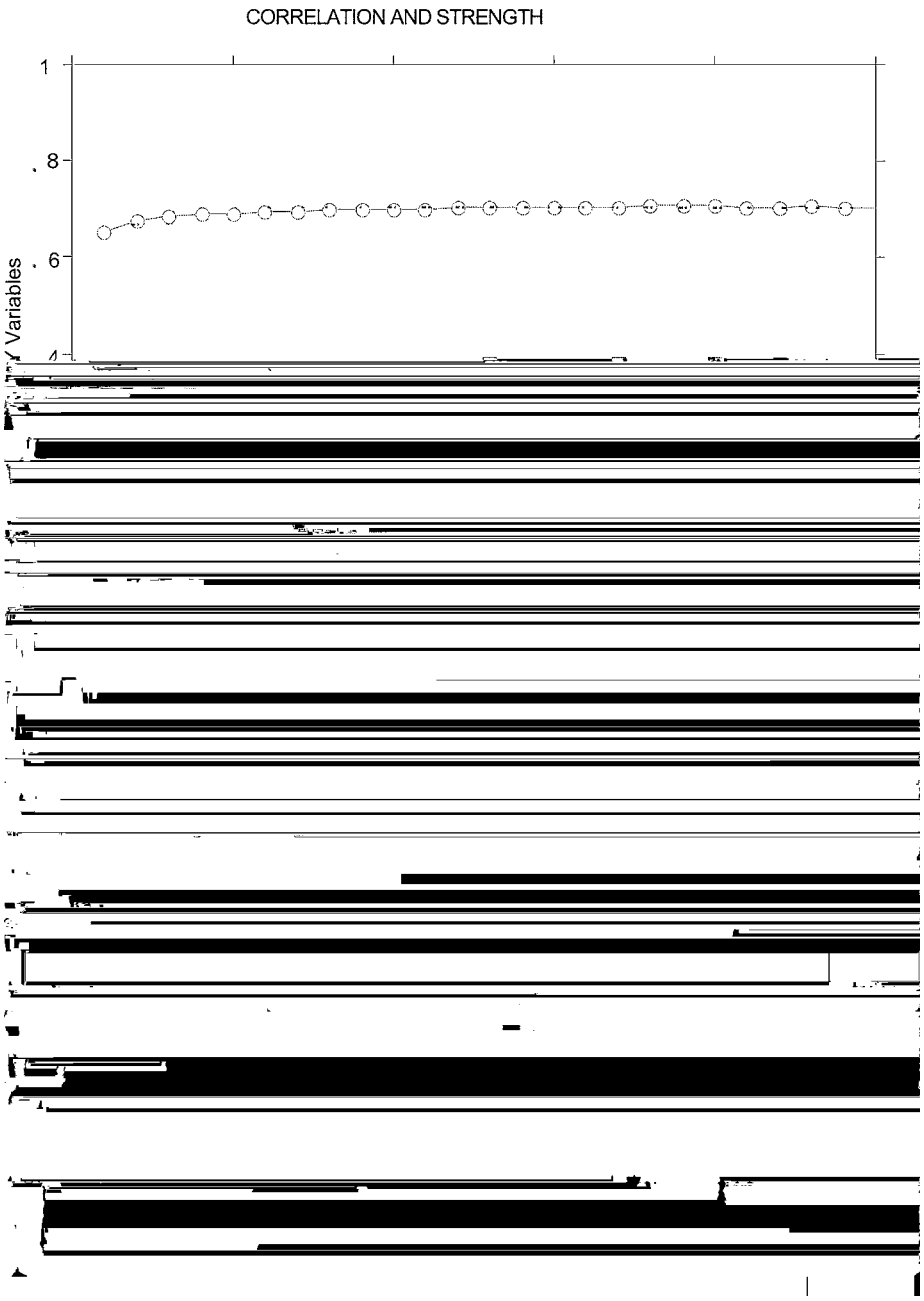

Figure 3. Effect of number of features on satellite data.

da a e . Bo h he correla ion and rengh ho a mall b read increa e. The error ra e ho a ligh decrea e. We conjec re ha in h larger and more comple da a e , he rengh com in e o increa e longer before i pla ea o .

O r re h indica e ha be er (lo er generali a ion error) random fore ha e lo er correla ion be een cla i er and higher rengh. The randomne ed in ree con r c ion ha o aim for lo correla ion $\bar{\rho}$ hile main aining rea onable rengh. Thi concl ion ha been him ed a in pre io o ork. Die erich (1998) ha mea re of di per ion of an en emble and no e ha more acc ra e en emble ha e larger di per ion. Fre nd (per onal comm - nica ion) belie e ha one rea on h Adaboo o ork o ell i ha a each ep i rie o deco ple he ne cla i er from he c rren one. Ami e al. (1999) gi e an anal i o ho ha he Adaboo algori hm i aimed a keeping he co ariance be een cla i er mall.

7. Conjecture: Adaboost is a random forest

Vario cla i er can be modi ed o e bo h a raining e and a e of eigh on he raining e . Con ider he follo ing random fore : a large collec ion of K differen e of non-nega i e m-one eigh on he raining e i de ned. Deno e he eigh b $\mathbf{w}(1), \mathbf{w}(2), \ldots, \mathbf{w}(K)$. Corre ponding o he eigh are probabili ie $p(1), p(2), \ldots, p(K)$ o e m i one. Dra from he in eger $1, \ldots, K$ according o he e probabili ie . The o come i Θ. If $\Theta = k$ gro he cla i er $h(\mathbf{x}, \Theta)$ ing he raining e in h eigh $\mathbf{w}(k)$.

In i original er ion, Adaboo (Fre nd & Schapire, 1996) i a de ermini ic algori hm ha elec he eigh on he raining e for inp o he ne cla i er ba ed on he mi - cla i ca ion in he pre io cla i er . In o r e perimen , random fore ere prod ced a follo : Adaboo a r n 75 ime on a da a e prod cing e of non-nega i e m-one eigh $\mathbf{w}(1), \mathbf{w}(2), \ldots, \mathbf{w}(50)$ (he r 25 ere di carded). The probabili for he kh e of eigh i e propor ional o $Q(\mathbf{w}_k) = \log[(1 - error(k))/error(k)]$ here $error(k)$ i he $\mathbf{w}(k)$eigh ed raining e error of he kh cla i er. Then he fore i r n 250 ime .

Thi a repea ed 100 ime on a fe da a e , each ime lea ing o 10% a a e e and hen a eraging he e e error . On each da a e , he Adaboo error ra e a er clo e o he random fore error ra e. A pical re h i on he Wi con in Brea Cancer da a here Adaboo prod ced an a erage of 2.91% error and he random fore prod ced 2.94%.

In he Adaboo algori hm, $\mathbf{w}(k+1) = \phi(\mathbf{w}(k))$ here ϕ i a f nc ion de ermined b he ba e cla i er. Deno e he kh cla i er b $h(\mathbf{x}, \mathbf{w}_k)$. The o e of he kh cla i er i eigh ed b $Q(w_k)$ o he normali ed o e for cla j a $\mathbf{x}$ eq al

$$\sum_k I(h(\mathbf{x}, \mathbf{w}_k) = j) Q(\mathbf{w}_k) \Big/ \sum_k Q(\mathbf{w}_k). \tag{10}$$

For an f nc ion f de ned on he eigh pace, de ne he opera or $\mathbf{T} f(\mathbf{w}) = f(\phi(\mathbf{w}))$. We conjec re ha $\mathbf{T}$ i ergodic in h in ariam mea re $\pi(d\mathbf{w})$. Then (10) ill con erge o $E_{Q\pi}[I(h(\mathbf{x}, \mathbf{w}) = j)]$ here he di rib ion $Q\pi(d\mathbf{w}) = Q(\mathbf{w})\pi(d\mathbf{w}) / \int Q(\mathbf{v})\pi(d\mathbf{v})$. If hi conjec re i r e, hen Adaboo i eq i alen o a random fore here he eigh on he raining e are elec ed a random from he di rib ion $Q\pi$.

It is true that in the two class case adaboost does not overfit as more trees are added to the ensemble, an experimental fact that has been puzzling. There is some experimental evidence that Adaboost may overfit if run long enough and of time (Grove & Schuurman, 1998), but the runs were done using rather simple base classifiers and may not carry over to the case of trees as the base classifier. My experience running Adaboost on 1,000 trees on a number of data sets is that the test set error converges to an asymptotic value.

The strength of this conjecture does not solve the problem of how Adaboost selects the favorable distribution on the training sets that it does. Note that the distribution of the training sets will depend on the training set. In the final random forest, the distribution of the random vectors does not depend on the training set.

8. The effects of output noise

Dietterich (1998) showed that when a fraction of the output labels in the training set are randomly altered, the accuracy of Adaboost degenerates, while bagging and random split selection are more immune to the noise. Since some noise in the output is often present, robustness with respect to noise is a desirable property. Following Dietterich (1998) the following experiment was done in which changed about one in twenty class labels (injecting 5% noise).

For each data set in the experiment, 10% are randomly split off as a test set. Three forests are made on the remaining training sets. The first forest is run on the training set as is. The second forest is run on a noise version of the training set. The noise version is gotten by changing, at random, 5% of the class labels into an alternative class label chosen uniformly from the other labels.

This is repeated 50 times using Adaboost (deterministic version), Forest-RI and Forest-RC. The test set errors are averaged over the 50 repetitions and the percent increase due to the noise computed. In both random forests, used the number of features giving the lowest test set error in the Section 5 and 6 experiments. Because of the length of the run, only the 9 smaller data sets are used. Table 4 gives the increase in error rate due to noise.

Table 4. Increase in error rate due to noise (%).

Data set	Adaboost	Forest-RI	Forest-RC
Glass	1.6	.4	−.4
Breast cancer	43.2	1.8	11.1
Diabetes	6.8	1.7	2.8
Sonar	15.1	−6.6	4.2
Ionosphere	27.7	3.8	5.7
Soybean	26.9	3.2	8.5
Ecoli	7.5	7.9	7.8
Votes	48.9	6.3	4.6
Liver	10.3	−.2	4.8

Adaboo deriora e markedl with 5% noi e, while he random fore proced re generall ho mall change . The effec on Adaboo i c rio l da a e dependen, with he o m hicla da a e , gla and ecoli, alongwih diabe e , lea effec ed b he noi e. The Adaboo algori hm i era i el increa e he eigh on he in ance mo recen l mi cla i ed. In ance ha ing incorrec cla label will per i in being mi cla i ed. Then, Adaboo will concen ra e increa ing eigh on he e noi in ance and become warped. The random fore proced re do no concen ra e eigh on an b e of he in ance and he noi e effec i maller.

9. Data with many weak inputs

Da a e with man weak inp are becoming more common, i.e. in medical diagno i , doc men re rie al, e c. The common charac eri ic i no ingle inp or mall gro p of inp can di ing i h be ween he cla e . Thi pe of da a i dif c l for he al cla i er , ne ral ne and ree .

To ee if here i a po ibili ha Fore -RI me hod can ork, he follo ing 10 cla , 1,000 binar inp da a,a genera ed: (rnd i a niform random n mber, elec ed ane each ime i appear)

```
do j=1,10
do k=1,1000
p(j,k)=.2*rnd+.01
end do
end do

do j=1,10
do i=1, nin(400*rnd)    !nin=neare  in eger
k=nin(1000*rnd)
p(j,k)=p(j,k)+.4*rnd
end do
end do

do n=1,N
j=nin(10*rnd)
do m=1,1000
if (rnd<p(j,m) ) hen
 (m,n)=1
el e
 (m,n)=0
end if
 (n)=j       !  (n) i he cla  label of he m h e ample
end do
end do
```

Thi code genera e a e of probabili ie {p(j, m)} here *j* i he cla label and m i he inp n mber. Then he inp for a cla j e ample are a ring of M binar ariable in h he m h ariable ha ing probabili p(j, m) of being one.

For he raining e, N = 1,000. A 4,000 e ample e e a al o con r c ed ing he ame {p (j, k)}. E amina ion of he code ho ha each cla ha higher nderl ing probabili a cer ain loca ion . B he o al o er all cla e of he e loca ion i abo 2,000, o here i igni cam o erlap. A ming one kno all of he {p (j, k)}, he Ba e error ra e for he pa ic lar {p (j, m)} comp ed in o r r n i 1.0%.

Since he inp are independen of each o her, he Nai e Ba e cla i er, hich e ima e he {p (j,k)} from he raining da a i ppo edl op imal and ha an error ra e of 6.2%. Thi i no an endor emen of Nai e Ba e , ince i o ld be ea o crea e a dependence be een he inp hich o ld increa e he Nai e Ba e error ra e. I a ed h hi e ample beca e he Ba e error ra e and he Nai e Ba e error ra e are ea o comp e.

I an ed h a r n of Fore -RI h $F = 1$. I con erged er lo l and b 2,500 ie ra ion , hen i a opped, i had ill no con erged. The e e error a 10.7%. The reng h a .069 and he correla ion .012 h a c/ 2 ra io of 2.5. E en ho gh he reng h a lo , he almo ero correla ion mean ha e ere adding mall incremen of acc rac a he i era ion proceeded.

Clearl , ha a de ired a an increa e in reng h hile keeping he correla ion lo . Fore -RI a r n again ing $F = in (\log_2 M + 1) = 10$. The re l ere enco raging. I con erged af er 2,000 i era ion . The e e error a 3.0%. The reng h a .22, he correla ion .045 and c/ 2 = .91. Going h he rend, Fore -RI a r n h $F = 25$ and opped af er 2,000 i era ion . The e e error a 2.8%. S reng h a .28, correla ion .065 and c/ 2 = .83.

I ' ime ere ing ha Fore -RI co ld prod ce error ra e no far abo e he Ba e error ra e. The indi id al cla i er are eak. For $F = 1$, he a erage ree error ra e i 80%; for $F = 10$, i i 65%; and for $F = 25$, i i 60%. Fore eem o ha e he abili o ork h er eak cla i er a long a heir correla ion i lo . A compari on ing Adaboo a ried, b I can ge Adaboo o r n on hi da a beca e he ba e cla i er are oo eak.

10. Exploring the random forest mechanism

A fore of ree i impene rable a far a imple im erpre a ion of i mechani m go. In ome applica ion , anal i of medical e perimen for e ample, i i cri ical o nder and he im erac ion of ariable ha i pro iding he predic i e acc rac . A an on hi problem i made b ing im ernal o -of-bag e ima e , and eri ca ion b rer n ing onl elec ed ariable .

S ppo e here are M inp ariable . Af er each ree i con r c ed, he al e of he m h ariable in he o -of-bag e ample are randoml perm ed and he o -of-bag da a i r n do n he corre ponding ree. The cla i ca ion gi en for each $\mathbf{x}_n$ ha i o of bag i a ed. Thi i repea ed for $m = 1, 2, \ldots, M$. A he end of he r n, he pl ra li of o -of-bag cla o e for $\mathbf{x}_n$ h he m h ariable noi ed p i compared h he r e cla label of $\mathbf{x}_n$ o gi e a mi cla i ca ion ra e.

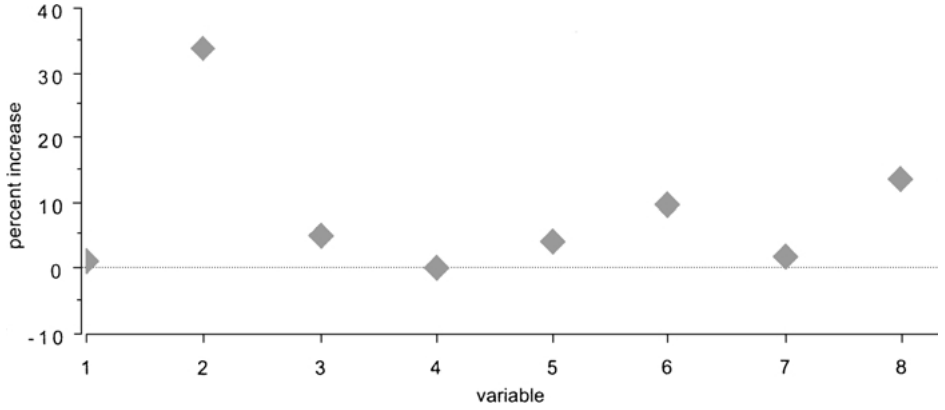

Figure 4. Mea re of ariable impor ance, diabe e da a.

The o p i he percen increa e in mi cla i ca ion ra e a compared o he o -of-bag ra e (wi h all ariable im ac). We ge he e e ima e b ing a ingle r n of a fore wi h 1,000 ree and no e e . The proced re i ill ra ed b e ample .

In he diabe e da a e , ing onl ingle ariable wi h $F = 1$, he ri e in error d e o he noi ing of ariable i gi en in g re 4

The econd ariable appear b far he mo impor an follo ed b ariable 8 and ariable 6. R ning he random fore in 100 repe i ion ing onl ariable 2 and lea ing o 10% each ime o mea re e e error ga e an error of 29.7%, compared wi h 23.1% ing all ariable . B wi hen ariable 8 i added, he error fall onl o 29.4%. When ariable 6 i added o ariable 2, he error fall o 26.4%.

The rea on ha ariable 6 eem impor an , e i no help once ariable 2 i en ered i a charac eri ic of ho dependen ariable affec predic ion error in random fore . Sa here are wo ariable x_1 and x_2 which are iden ical and carr igni can predic i e informa ion. Beca e each ge picked wi h abo he ame freq enc in a random fore , noi ing each epara el ill re l in he ame increa e in error ra e. B once x_1 i en ered a a predic i e ariable, ing x_2 in addi ion ill no prod ce an decrea e in error ra e. In he diabe e da a e , he 8 h ariable carrie ome of he ame informa ion a he econd. So i doe no add predic i e acc rac hen combined wi h he econd.

The rela i e magni de of ri e in error ra e are fairl able wi h re pec o he inp fea re ed. The e perimen abo e wa repea ed ing combina ion of hree inp wi h $F = 2$. The re l are in g re 5.

Ano her in ere ing e ample i he oring da a. Thi ha 435 e ample corre ponding o 435 Congre men and 16 ariable re ec ing heir e -no o e on 16 i e . The cla ariable i Rep blican or Democra . To ee hich i e were mo impor an, we again ran he noi ing ariable program genera ing 1,000 ree . The lo e error ra e on he original da a wa go en ing ingle inp wi h $F = 5$, o he e parame er were ed in he r n. The re l in g re 6.

Variable 4 and o he error riple if ariable 4 i noi ed. We reran hi da a e ing onl ariable 4. The e e error i 4.3%, abo he ame a if all ariable were ed. The

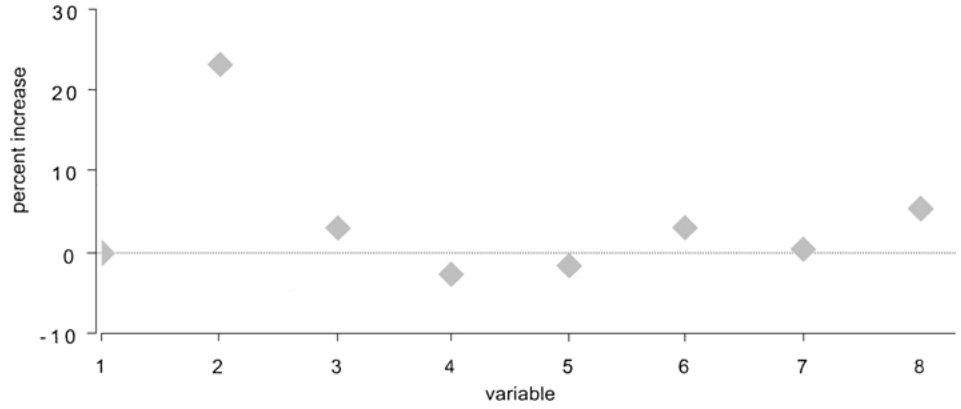

Figure 5. Mea re of ariable impor ance-diabe e da a.

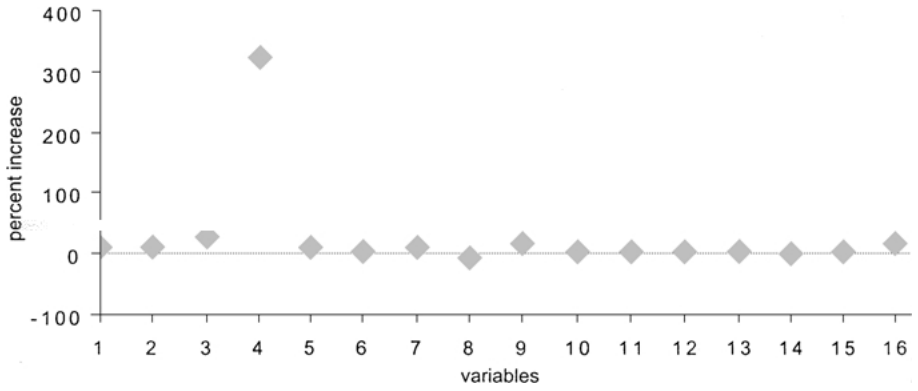

Figure 6. Mea re of ariable impor ance o e da a.

o e on 4 epara e Rep blican from Democra almo a well a he o e on 4 combined ih he o e on all o her 15 i e .

The approach gi en in hi ec ion i onl a beginning. More re earch ill be nece ar o nder and ho o gi e a more comple e pic re.

11. Random forests for regression

Random fore for regre ion are formed b gro ing ree depending on a random ec or Θ ch ha he ree predic or $h(\mathbf{x}, \Theta)$ ake on n merical al e a oppo ed o cla label . The o p al e are n merical and e a me ha he raining e i independen l dra n from he di rib ion of he random ec or Y, $\mathbf{X}$. The mean- q ared generali a ion error for an n merical predic or $h(\mathbf{x})$ i

$$E_{\mathbf{X},Y}(Y - h(\mathbf{X}))^2 \tag{11}$$

The random fore predic or i formed b aking he a erage o er k of he ree $\{h(\mathbf{x}, \Theta_k)\}$. Similarl o he cla i ca ion ca e, he follo ing hold :

Theorem 11.1. *As the number of trees in the forest goes to infinity, almost surely,*

$$E_{\mathbf{X},Y}(Y - av_k h(\mathbf{X}, \Theta_k))^2 \to E_{\mathbf{X},Y}(Y - E_\Theta h(\mathbf{X}, \Theta))^2. \tag{12}$$

Proof: ee Appendi I. □

Deno e he righ hand ide of (12) a PE*(fore), he generali a ion error of he fore . De ne he a erage generali a ion error of a ree a :

$$\text{PE}^*(\text{ree}) = E_\Theta\, E_{\mathbf{X},Y}(Y - h(\mathbf{X}, \Theta))^2$$

Theorem 11.2. *Assume that for all* Θ, $EY = E_{\mathbf{X}} h(\mathbf{X}, \Theta)$. *Then*

$$PE^*(forest) \le \bar{\rho} PE^*(tree)$$

where $\bar{\rho}$ *is the weighted correlation between the residuals* $Y - h(\mathbf{X}, \Theta)$ *and* $Y - h(X, \Theta')$ *where* Θ, Θ' *are independent.*

Proof:

$$\begin{aligned} PE^*(forest) &= E_{\mathbf{X},Y}[E_\Theta(Y - h(\mathbf{X}, \Theta)]^2 \\ &= E_\Theta E_{\Theta'} E_{\mathbf{X},Y}(Y - h(\mathbf{X}, \Theta))(Y - h(\mathbf{X}, \Theta')) \end{aligned} \tag{13}$$

The erm on he righ in (13) i a co ariance and can be e ri en a :

$$E_\Theta E_{\Theta'}(\rho(\Theta, \Theta') sd(\Theta) sd(\Theta'))$$

here $sd(\Theta) = \sqrt{E_{\mathbf{X},Y}(Y - h(\mathbf{X}, \Theta))^2}$. De ne he eighed correla ion a :

$$\bar{\rho} = E_\Theta E_{\Theta'}(\rho(\Theta, \Theta') sd(\Theta) sd(\Theta'))/(E_\Theta sd(\Theta))^2 \tag{14}$$

Then

$$PE^*(forest) = \bar{\rho}(E_\Theta sd(\Theta))^2 \le \bar{\rho} PE^*(tree). \qquad \square$$

Theorem (11.2) pinpoi he req iremen for acc ra e regre ion fore , lo correla ion be een re id al and lo error ree . The random fore decrea e he a erage error of he ree emplo ed b he fac or $\bar{\rho}$. The randomi a ion emplo ed need o aim a lo correla ion.

12. Empirical results in regression

In regre ion fore , e random fea re elec ion on op of bagging. Therefore, e can e he moni oring pro ided b o -of-bag e ima ion o gi e e ima e of PE*(fore), PE*(ree) and $\bar{\rho}$. The e are deri ed imilarl o he e ima e in cla i ca ion. Thro gho , fea re formed b a random linear m of o inp are ed. We commen la er on ho man of he e fea re o e o determine he pli a each node. The more fea re ed, he lo er PE*(ree) b he higher $\bar{\rho}$. In o r empirical d he follo ing da a e are ed, ee Table 5.

Of he e da a e , he Bo on Ho ing, Abalone and Ser o are a ailable a he UCI repo i or . The Robo Arm da a a pro ided b Michael Jordan. The la hree da a e are m he ic. The origina ed in Friedman (1991) and are al o de cribed in Breiman (1998b). The e are he ame da a e ed o compare adap i e bagging o bagging (ee Breiman, 1999), e cep ha one m he ic da a e, (Peak20), hich a fo nd anomalo bo h b o her re earcher and m elf, i elimina ed.

The r hree da a e li ed are modera e in i e and e e error a e ima ed b lea ing o a random 10% of he in ance , r nning on he remaining 90% and ing he lef -o 10% a a e e . Thi a repea ed 100 ime and he e e error a eraged. The abalone da a e i larger i h 4,177 in ance and 8 inp ariable . It originall came i h 25% of he in ance e a ide a a e e . We ran hi da a e lea ing o a randoml elec ed 25% of he in ance o e a a e e , repea ed hi 10 ime and a eraged.

Table 6 gi e he e e mean- q ared error for bagging, adap i e bagging and he random fore . The e ere all r n ing 25 fea re , each a random linear combina ion of o randoml elec ed inp , o pli each node, each fea re a random combina ion of o inp . All r n i h all da a e , combined 100 ree . In all da a e , he r le don pli if he node i e i <5 a enforced.

An in ere ing difference be een regre ion and cla i ca ion i ha he correla ion increa e q i e lo l a he n mber of fea re ed increa e . The major effec i he decrea e in PE*(ree). Therefore, a rela i el large n mber of fea re are req ired o red ce PE*(ree) and ge near op imal e e error.

Table 5. Da a e mmar .

Da a e	Nr. inp	#Training	#Te
Bo on Ho ing	12	506	10%
O one	8	330	10%
Ser o	4	167	10%
Abalone	8	4177	25%
Robo Arm	12	15,000	5000
Friedman#1	10	200	2000
Friedman#2	4	200	2000
Friedman#3	4	200	2000

Table 6. Mean- q ared e e error.

Da a e	Bagging	Adap . bag	Fore
Bo on Ho ing	11.4	9.7	10.2
O one	17.8	17.8	16.3
Ser o $\times 10-2$	24.5	25.1	24.6
Abalone	4.9	4.9	4.6
Robo Arm $\times 10-2$	4.7	2.8	4.2
Friedman #1	6.3	4.1	5.7
Friedman #2 $\times 10+3$	21.5	21.5	19.6
Friedman #3 $\times 10-3$	24.8	24.8	21.6

The re h ho n in Table 6 are mi ed. Random fore -random fea re i a a be er han bagging. In da a e for hich adap i e bagging gi e harp decrea e in error, he decrea e prod ced b fore are no a prono nced. In da a e in hich adap i e bagging gi e no impro emen o er bagging, fore prod ce impro emen .

For he ame n mber of inp combined, o er a ide range, he error doe no change m ch i h he n mber of fea re . If he n mber ed i oo mall, PE*(ree) become oo large and he error goe p. If he n mber ed i oo large, he correla ion goe p and he error again increa e . The in-be een range i all large. In hi range, a he n mber of fea re goe p, he correla ion increa e , b PE*(ree) compen a e b decrea ing.

Table 7 gi e he e e error , he o -of-bag error e ima e , and he OB e ima e for PE*(ree) and he correla ion.

A e pec ed, he OB Error e ima e are con i em l high. I i lo in he robo arm da a, b I belie e ha hi i an an ifac ca ed b epara e raining and e e , here he e e ma ha e a ligh l higher error ra e han he raining e .

A an e perimen , I rned off he bagging and replaced i b randomi ing o p (Breiman, 1998b). In hi proced re, mean- ero Ga ian noi e i added o each of he o p . The andard de ia ion of he noi e i e eq al o he andard de ia ion of he

Table 7. Error and OB e ima e .

Da a Se	Te error	OB error	PE*(ree)	Cor.
Bo on Ho ing	10.2	11.6	26.3	.45
O one	16.3	17.6	32.5	.55
Ser o $\times 10-2$	24.6	27.9	56.4	.56
Abalone	4.6	4.6	8.3	.56
Robo Arm $\times 10-2$	4.2	3.7	9.1	.41
Friedman #1	5.7	6.3	15.3	.41
Friedman #2 $\times 10+3$	19.6	20.4	40.7	.51
Friedman #3 $\times 10-3$	21.6	22.9	48.3	.49

Table 8. Mean- q ared e y e error.

Da a e	Wi h bagging	Wi h Noi e
Bo on Ho ing	10.2	9.1
O one	17.8	16.3
Ser o $\times 10-2$	24.6	23.2
Abalone	4.6	4.7
Robo Arm $\times 10-2$	4.2	3.9
Friedman #1	5.7	5.1
Friedman #2 $\times$ 10 + 3	19.6	20.4
Friedman #3 $\times 10-3$	21.6	19.8

o p . Similar o he bagging e perimem , ree con r c ion a done ing 25 fea re , each a random linear combina ion of wo randoml elec ed inp , o pli each node. The re l are gi en in Table 8.

The error ra e on he r wo da a e are he lo e o da e. O erall, adding o p noi e ork wi h random fea re elec ion be er han bagging. Thi i ill ra i e of he e ibili of he random fore e ing ario combina ion of randomne can be added o ee ha ork he be .

13. Remarks and conclusions

Random fore are an effec i e ool in predic ion. Beca e of he La of Large N mber he do no o er . Injec ing he righ kind of randomne make hem acc ra e cla i er and regre or . F rthermore, he frame ork in erm of reng h of he indi id al predic or and heir correla ion gi e in igh in o he abili of he random fore o predic . U ing o -of-bag e ima ion make concre e he o her i e heore ical al e of reng h and correla ion.

For a hile, he con en ional hinking a ha fore co ld no compe e wi h arcing pe algori hm in erm of acc rac . O r re l di pel hi belief, b lead o in ere ing q e ion . Boo ing and arcing algori hm ha e he abili o red ce bia a ell a ariance (Schapire e al., 1998). The adap i e bagging algori hm in regre ion (Breiman, 1999) a de igned o red ce bia and opera e effec i el in cla i ca ion a ell a in regre ion. B , like arcing, i al o change he raining e a in progre e .

Fore gi e re l compe i i e wi h boo ing and adap i e bagging, e do no progre - i el change he raining e . Their acc rac indica e ha he ac o red ce bia . The mechani m for hi i no ob io . Random fore ma al o be ie ed a a Ba e ian pro- ced re. Al ho gh I do b ha hi i a fr i f l line of e plora ion, if i co ld e plain he bia red c ion, I migh become more of a Ba e ian.

Random inp and random fea re prod ce good re l in cla i ca ion le o in regre ion. The onl pe of randomne ed in hi d i bagging and random fea re . I ma ell be ha o her pe of injec ed randomne gi e be er re l . For in ance, one of he referee ha gge ed e of random Boolean combina ion of fea re .

An almo ob io q e ion i he her gain in acc rac can be go en b combining random fea re i h boo ing. For he larger da a e , i eem ha igni cam l lo er error ra e are po ible. On ome r n , e go error a lo a 5.1% on he ip-code da a, 2.2% on he le er da a and 7.9% on he a elli e da a. The impro emen a le on he maller da a e . More ork i needed on hi ; b i doe gge ha differen injec ion of randomne can prod ce be er re l .

A recen paper (Breiman, 2000) ho (ha in di rib ion pace for o cla problem , random fore are eq i alen o a kernel ac ing on he r e margin. Arg men are gi en ha randomne (lo correla ion) enforce he mme r of he kernel hile reng h enhance a de irable ke ne a abr p c r ed bo ndarie . Hopef ll , hi hed ligh on he d al role of correla ion and reng h. The heore ical frame ork gi en b Kleinberg (2000) for S ocha ic Di crimina ion ma al o help nder anding.

Appendix I: Almost sure convergence

Proof of theorem 1.2: I f ce o ho (ha here i a e of probabili ero C on he eq ence pace $\Theta_1, \Theta_2, \ldots$ ch ha o ide of C, for all $\mathbf{x}$,

$$\frac{1}{N}\sum_{n=1}^{N} I\left(h(\Theta_n, \mathbf{x}) = j\right) \to P_\Theta(h(\Theta, \mathbf{x}) = j).$$

For a ed raining e and ed Θ, he e of all $\mathbf{x}$ ch ha $h(\Theta, \mathbf{x}) = j$ i a nion of h per-rec angle . For all $h(\Theta, \mathbf{x})$ here i onl a ni e n mber K of ch nion of h per-rec angle , deno ed b $S_1, \ldots, S_K$. De ne $\varphi(\Theta) = k$ if $\{\mathbf{x} : h(\Theta, \mathbf{x}) = j\} = S_k$. Le N_k be he n mber of ime ha $\varphi(\Theta_n) = k$ in he r N rial . Then

$$\frac{1}{N}\sum_{n=1}^{N} I\left(h(\Theta_n, \mathbf{x}) = j\right) = \frac{1}{N}\sum_{k} N_k I\left(\mathbf{x} \in S_k\right)$$

B he La of Large N mber ,

$$N_k = \frac{1}{N}\sum_{n=1}^{N} I\left(\varphi(\Theta_n) = k\right)$$

con erge a. o $P_\Theta(\varphi(\Theta) = k)$. Taking nion of all he e on hich con ergence doe no occ r for ome al e of k gi e a e C of ero probabili ch ha o ide of C,

$$\frac{1}{N}\sum_{n=1}^{N} I\left(h(\Theta_n, \mathbf{x}) = j\right) \to \sum_{k} P_\Theta(\varphi(\Theta) = k) I\left(\mathbf{x} \in S_k\right).$$

The righ hand ide i $P_\Theta(h(\Theta, \mathbf{x}) = j$. □

Proof of theorem 9.1: There are at most one of the pre-rectangles $R_1, \ldots, R_K$, such that if $\bar{y}_k$ is the average of the training y-values for all training inputs vectors in R_k then $h(\Theta, \mathbf{x})$ has one of the values $I(\mathbf{x} \in S_k)\bar{y}_k$. The rest of the proof parallels that of Theorem 1.2. □

Appendix II: Out-of bag estimates for strength and correlation

At the end of a combination run, let

$$Q(\mathbf{x}, j) = \sum_k I(h(\mathbf{x}, \Theta_k) = j; (y, \mathbf{x}) \notin T_{k,B}) \Big/ \sum_k I((y, \mathbf{x}) \notin T_{k,B}).$$

Then, $Q(\mathbf{x}, j)$ is the out-of-bag proportion of times case $(y, \mathbf{x})$ for class j, and is an estimate for $P_\Theta(h(\mathbf{x}, \Theta) = j)$. From Definition 2.1 the strength is the expectation of

$$P_\Theta(h(\mathbf{x}, \Theta) = y) - \max_{j \neq \mathbf{y}} P_\Theta(h(\mathbf{x}, \Theta) = j)$$

Substituting $Q(\mathbf{x}, j)$, $Q(\mathbf{x}, y)$ for $P_\Theta(h(\mathbf{x}, \Theta) = j)$, $P_\Theta(h(\mathbf{x}, \Theta) = y)$ in this last expression and taking the average over the training gives the strength estimate.

From Eq. (7),

$$\bar{\rho} = \text{var}(mr)/(E_\Theta sd(\Theta))^2.$$

The variance of mr is

$$E_{\mathbf{X},Y}[P_\Theta(h(\mathbf{x}, \Theta) = y) - \max_{j \neq \mathbf{y}} P_\Theta(h(\mathbf{x}, \Theta) = j)]^2 - s^2 \tag{A1}$$

where s is the strength. Replacing the first term in (A1) by the average over the training set of

$$(Q(\mathbf{x}, y) - \max_{j \neq \mathbf{y}} Q(\mathbf{x}, j))^2$$

and s by the out-of-bag estimate of s gives the estimate of $\text{var}(mr)$. The standard deviation is given by

$$sd(\Theta) = [p_1 + p_2 + (p_1 - p_2)^2]^{1/2} \tag{A2}$$

where

$$p_1 = E_{\mathbf{X},Y}(h(\mathbf{X}, \Theta) = Y)$$
$$p_2 = E_{\mathbf{X},Y}(h(\mathbf{X}, \Theta) = \hat{j}(\mathbf{X}, Y))$$

After the kth classifier is constructed, $Q(\mathbf{x}, j)$ is computed, and used to compute $\hat{j}(\mathbf{x}, y)$ for each example in the training set. Then, let p_1 be the average over all $(y, \mathbf{x})$ in the training set but not in the kth bagged training set of $I(h(\mathbf{x}, \Theta_k) = y)$. Then p_2 is the similar average of $I(h(\mathbf{x}, \Theta_k) = \hat{j}(\mathbf{x}, y))$. Subtracting this estimate from the estimate in (A2) gives an estimate of $sd(\Theta_k)$. Averaging the $sd(\Theta_k)$ over all k gives the final estimate of $sd(\Theta)$.

References

Amit, Y. & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation, 9*, 1545–1588.

Amit, Y., Blanchard, G., & Wilder, K. (1999). Multiple randomized classifiers: MRCL Technical Report, Department of Statistics, University of Chicago.

Bauer, E. & Kohavi, R. (1999). An empirical comparison of voting classification algorithms. *Machine Learning, 36*(1/2), 105–139.

Breiman, L. (1996a). Bagging predictors. *Machine Learning 26*(2), 123–140.

Breiman, L. (1996b). Out-of-bag estimation, ftp.stat.berkeley.edu/pub/users/breiman/OOBestimation.ps

Breiman, L. (1998a). Arcing classifier (discussion paper). *Annals of Statistics, 26*, 801–824.

Breiman. L. (1998b). Randomizing outputs to increase prediction accuracy. Technical Report 518, May 1, 1998, Statistics Department, UCB (in press, Machine Learning).

Breiman, L. 1999. Using adaptive bagging to debias regressions. Technical Report 547, Statistics Dept. UCB.

Breiman, L. 2000. Some infinity theory for predictor ensembles. Technical Report 579, Statistics Dept. UCB.

Dietterich, T. (1998). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization, *Machine Learning*, 1–22.

Freund, Y. & Schapire, R. (1996). Experiments with a new boosting algorithm, *Machine Learning*: *Proceedings of the Thirteenth International Conference*, 148–156.

Grove, A. & Schuurmans, D. (1998). Boosting in the limit: Maximizing the margin of learned ensembles. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 20*(8), 832–844.

Kleinberg, E. (2000). On the algorithmic implementation of stochastic discrimination. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 22*(5), 473–490.

Schapire, R., Freund, Y., Bartlett, P., & Lee, W. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics, 26*(5), 1651–1686.

Tibshirani, R. (1996). Bias, variance, and prediction error for classification rules. Technical Report, Statistics Department, University of Toronto.

Wolpert, D. H. & Macready, W. G. (1997). An efficient method to estimate Bagging's generalization error (in press, Machine Learning).

Received November 30, 1999
Revised April 11, 2001
Accepted April 11, 2001
Final manuscript April 11, 2001