What Size Test Set Gives Good Error Rate Estimates?

Isabelle Guyon, John Makhoul, *Fellow, IEEE*, Richard Schwartz, and Vladimir Vapnik

Abstract—We address the problem of determining what size test set guarantees statistically significant results in a character recognition task, as a function of the expected error rate. We provide a statistical analysis showing that if, for example, the expected character error rate is around 1 percent, then, with a test set of at least 10,000 statistically independent handwritten characters (which could be obtained by taking 100 characters from each of 100 different writers), we guarantee, with 95 percent confidence, that: (1) The expected value of the character error rate is not worse than 1.25 E, where E is the empirical character error rate of the best recognizer, calculated on the test set; and (2) a difference of 0.3 E between the error rates of two recognizers is significant. We developed this framework with character recognition applications in mind, but it applies as well to speech recognition and to other pattern recognition problems.

Index Terms—Pattern recognition, test set, test set size, benchmark, hypothesis testing, designed experiment, statistical significance, estimation, guaranteed estimators, recognition error.

1 Introduction

The problem often arises when organizing benchmarks in pattern recognition to determine what size test set will give statistically significant results. This is a chicken and egg problem, since before getting the recognizer performance, it is not possible to determine the statistical significance. Nevertheless, since approximate values of the error rates of particular recognizers on similar tasks are known, it is possible to estimate what reasonable size a test set should have. In this paper, we use fairly straightforward statistical arguments [1] to address that problem. The method has been designed to help in preparing the data for the first UNIPEN benchmark [2], but the results are fairly general and a broader applicability is expected.

We tackle the problem from the point of view of the benchmark organizer. Thus, our approach differs from the classical "hypothesis testing" framework (see, e.g., [1]) in that we do not test the statistical significance of the result of an actual experiment. Rather, we seek bounds on the minimum number of test examples that guarantee our future benchmark to provide: a good estimate of the state-of-the-art error rate on the target task and good confidence that one system is better than another, for a relatively small difference in their error rates.

1) We introduce the principle of our method and the notion of *guaranteed estimators*.
2) We estimate test set sizes, assuming that the errors are independently and identically distributed.
3) We introduce the problem of "correlations" between errors due, for instance, to having many consecutive examples provided by the same writer. We generalize the results to the case of multiple factors of correlation, including: recording conditions and linguistic material.
4) We treat the problem of the statistical significance of the difference in performance of two recognizers.
5) We summarize the practical aspects for determining the number of examples necessary to obtain statistical significance and analyze examples.
6) We suggest some statistical tests to be performed after the benchmark to verify the quality of the results.

The reader interested in only practical aspects of the results can go directly to Section 6.

2 Principle of the Method: Guaranteed Estimators

2.1 Punctual Estimators and Guaranteed Estimators

The problem addressed in a pattern recognition benchmark is to calculate and compare error rates of various recognizers. The error rate p of a given recognizer is estimated by computing the average error $\hat{p}$ over a finite number n of test examples or patterns. Let x_i, $i = 1, ..., n$, represent the recognition results for the test patterns, i.e., $x_i = 1$ if there is a recognition error for pattern i, and $x_i = 0$ otherwise. The average error rate, then, is computed as

$$\hat{p} = \frac{1}{n}\sum_{i}^{n} x_i \,. \tag{1}$$

Patterns are assumed to be drawn randomly and independently from a source of patterns. For a particular

- *I. Guyon is an independent consultant working at 955 Creston Road, Berkeley, CA 94708. E-mail: isabelle@clopinet.com.*
- *V. Vapnik is with AT&T Labs, Red Bank, NJ 07701. E-mail: vlad@research.att.com.*
- *J. Makhoul and R. Schwartz are with BBN Systems and Technologies, Cambridge, MA 02138. E-mail: {makhoul, schwartz}@bbn.com.*

Manuscript received 11 Dec. 1995; revised 25 Aug. 1997. Recommended for acceptance by J.J. Hull.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 105690.

recognizer, the failure or success of recognition of the *i*th pattern is the realization x_i of a random variable X_i. The random variable

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2}$$

is a *punctual estimator* of the mean, the expected value of which is p. The average error rate $\hat{p}$ is a realization of $\overline{X}$.

For pattern recognition benchmarks we are also interested in confidence intervals. Two scenarios are possible here. With a certain confidence $(1 - \alpha)$, $0 \le \alpha \le 1$, we want the expected value of the error rate p to be either within a certain range:

$$\hat{p} - \varepsilon(n,\alpha) < p < \hat{p} + \varepsilon(n,\alpha) \tag{3}$$

(two-sided risk), or simply not to exceed a certain value:

$$p < \hat{p} + \varepsilon(n,\alpha) \tag{4}$$

(one-sided risk). In this paper, we use the one-sided risk because it is not of concern to us if the expected value of the error rate is better than what we estimate.

The random variable of which $\hat{p} + \varepsilon(n, \alpha)$ is a realization is a *guaranteed estimator* of the mean. We are guaranteed, with risk α of being wrong, that the mean does not exceed $\hat{p} + \varepsilon(n, \alpha)$:

$$\text{Prob}\big(p \ge \hat{p} + \varepsilon(n,\alpha)\big) \le \alpha. \tag{5}$$

2.2 Methods for Obtaining Estimators

Punctual estimators are often obtained by the Maximum Likelihood (ML) method. For instance, the estimator $\overline{X} = (1/n)\sum_{i=1}^{n} X_i$ is the ML estimator of the mean for the Gaussian distribution. It is consistent (its realizations converge to the mean for an infinite amount of examples) and unbiased (its expected value is equal to the mean).

Guaranteed estimators are obtained either from the properties of the underlying probability distribution, if it is known, or from distribution-independent bounds. One of the most well-known distribution-independent bounds is the Chebychev inequality (see, e.g., [1]):

$$\text{Prob}\left(\left|p - \hat{p}\right| \ge \frac{\sigma}{\sqrt{\alpha n}}\right) \le \alpha, \tag{6}$$

or, for the one-sided version:

$$\text{Prob}\left(p - \hat{p} \ge \frac{\sigma}{\sqrt{2\alpha n}}\right) \le \alpha, \tag{7}$$

where σ^2 is the variance of X_i, estimated, for instance, as:[1]

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{p} - x_i\right)^2, \tag{8}$$

Other tighter bounds have been proposed more recently by Chernoff [3], Hoeffding [4], and others for the binomial distribution. Those bounds are tighter than Chebychev's inequality, but Chebychev's inequality is distribution independent.

1. The denominator $(n - 1)$ can be approximated by n for large values of n. It accounts for the fact that p is not known and is also estimated from data, which removes one "degree of freedom."

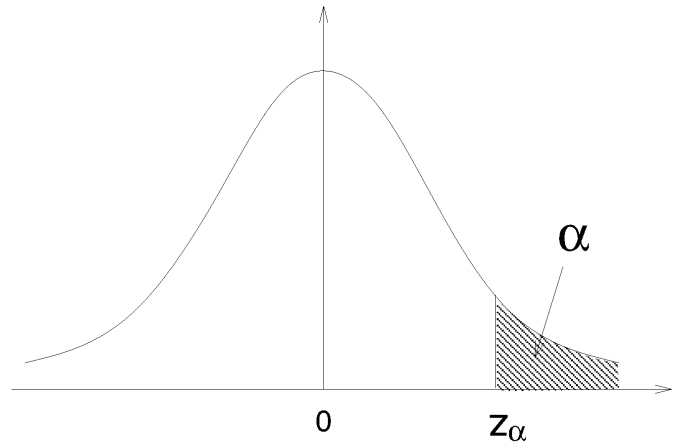

Fig. 1. One sided risk: With probability $(1 - \alpha)$, $z < z_\alpha$.

Better bounds are obtained if more is known about the probability distribution. In particular, assume that X_i is distributed according to the Normal law (Gaussian distribution) of mean $\mu = p$ and variance σ^2, and with probability function:

$$\rho_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \tag{9}$$

The random variable $\overline{X} = \sum_{i=1}^{n} X_i$ is distributed according to the Normal law of mean np and of variance the sum of the variances: $n\sigma^2$ (assuming the X_i are independent). Thus the random variable $\overline{X} = (1/n)\sum_{i=1}^{n} X_i$ is distributed according to the Normal law of mean p and of variance σ^2/n. Consequently, the random variable:

$$Z = \frac{p - \overline{X}}{\sigma/\sqrt{n}} \tag{10}$$

obeys the standardized Normal law (with mean 0 and variance 1).[2] The distribution of this law is tabulated, which allows us to determine the threshold z_α under which we find all realizations of Z with probability $(1 - \alpha)$ (Fig. 1). The bound of interest, then, is:

$$\text{Prob}\left(p - \hat{p} \ge \frac{z_\alpha \sigma}{\sqrt{n}}\right) \le \alpha, \tag{11}$$

where $\hat{p}$ is a realization of $\overline{X}$.

The standardized Normal law distribution table provides values of z_α for various values of one-sided risk α (see, e.g., [1]). A relatively good approximation of z_α in this range of values is given by:

$$z_\alpha \simeq \sqrt{-\ln \alpha}, \tag{12}$$

where ln is the Neperian logarithm (see Table 1). This approximation is convenient since it provides us with a functional relation between α and z_α which will prove to be useful in our calculations.

2. When the variance is not known and has to be estimated from data as well as the mean, Z obeys Student's law with $(n - 1)$ degrees of freedom. For values of n sufficiently large $(n > 30)$, the Normal law is a very good approximation to Student's law.

TABLE 1
VALUES OF z_α FOR ONE-SIDED RISK AND RELATED COEFFICIENTS APPEARING IN OTHER BOUNDS

α	z_α	z_α^2	$\sqrt{-\ln \alpha}$	$-\ln \alpha$	$\sqrt{-2 \ln \alpha}$	$-2 \ln \alpha$
0.01	2.33	5.29	2.15	4.61	3.03	9.21
0.05	1.65	2.72	1.73	3.00	2.45	5.99
0.10	1.28	1.64	1.52	2.30	2.15	4.60

2.3 Number of Test Examples Needed

Before the benchmark, *guaranteed estimators* (inequality (4)) are used to determine the number of test examples needed to guarantee a certain margin of error $\varepsilon(n, \alpha)$ (e.g., $\varepsilon(n, \alpha) = z_\alpha \sigma / \sqrt{n}$ for the Normal law).

In this paper, we fix $\varepsilon(n, \alpha)$ to be a given fraction of p:

$$\varepsilon(n, \alpha) = \beta p \tag{13}$$

and we solve (13) for n to obtain the desired number of test examples.

The values of p and σ which are necessary to determine n are generally unknown. Thus our estimate of n will depend on the hypotheses we make for p and σ. These hypotheses are based on the results of other similar benchmarks and/or on human performance. After the benchmark, actual values of $\hat{p}$ and $\hat{\sigma}$ are computed and *guaranteed estimators* can be used again to verify the statistical significance of the results (hypothesis testing, see Section 7).

3 TEST SET SIZE NEEDED FOR I.I.D. ERRORS

3.1 Recognition Errors as Bernoulli Trials

In many benchmarks, the errors on the test examples are not independently and identically distributed (i.i.d). In particular, for speech and handwriting recognition, speaker/writer-independent tasks are usually tested with data containing long sequences of examples from each of a number of speakers/writers (see Section 4). There may be also error correlations introduced by the recognizer itself if, for instance, use is made of a language model. In the present section, we consider the simple case of i.i.d. errors, an illustration of which could be a speaker/writer-dependent isolated word recognition task, using a specific vocabulary distribution and specific recording conditions.

Consider a source of i.i.d. data which are drawn according to a certain probability distribution $P(pattern, class) = P(pattern)\, P(class \mid pattern)$ and a recognizer which recognizes those data independently of each other with a probability of error p. The ensemble {data source, recognizer} is a source of binary events: 1 for error and 0 for no error, with probability p of drawing a 1 and $(1 - p)$ of drawing a 0 (Fig. 2). Such a random process is known under various names, including "random walk" and "Bernoulli trials." The random variable K counting the number of errors in n trials is distributed according to the binomial distribution:

$$\rho_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k}, \tag{14}$$

of mean np and variance $np(1 - p)$.

For a test set size of n examples, the following is an estimate of p:

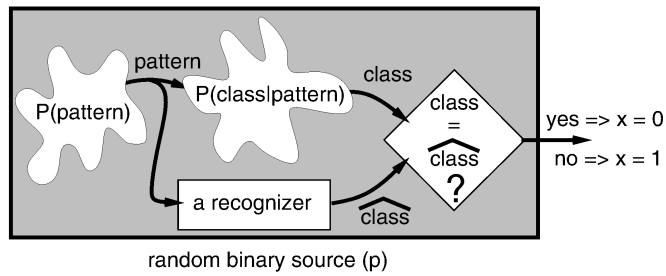

Fig. 2. Recognition process of i.i.d. data: patterns are, for instance, handwritten characters and class labels are, for instance, "0," "1," ..., "a," "b,"... The ensemble {data source, recognizer} is a random binary source which produces 1 with probability p and 0 with probability $(1 - p)$, where p is the expected value of the error rate of the recognizer.

$$\hat{p} = \frac{k}{n}, \tag{15}$$

where k is the number of errors. The expected value of the error rate is p and $\hat{p}$ is the *empirical* value of the error rate estimated on the test set.

We are seeking a *guaranteed estimator* which provides the guarantee that, with probability $(1 - \alpha)$, p is not larger than $\hat{p}$ plus a certain error $\varepsilon(n, \alpha)$:

$$\text{Prob}\big(p \geq \hat{p} + \varepsilon(n, \alpha)\big) = \sum_{np - k \geq \varepsilon n} \rho_{n,p}(k) \leq \alpha. \tag{16}$$

If we express $\varepsilon(n, \alpha)$ as a small fraction β of p, then (16) becomes:

$$\sum_{k \leq (1-\beta)np} \rho_{n,p}(k) \leq \alpha. \tag{17}$$

We are interested in solving this equation for n but, unfortunately, there is no analytical solution. Furthermore, a numerical solution is tedious. To simplify matters, we approximate the binomial law by the Normal law (probability function (9)) of mean np and of variance $np(1 - p)$.

With this approximation,

$$Z = \frac{p - \frac{K}{n}}{\sqrt{\frac{p(1-p)}{n}}}. \tag{18}$$

obeys the standardized Normal law (with mean 0 and variance 1).

Similarly, (11) reduces to:

$$\text{Prob}\left(p - \hat{p} \geq z_\alpha \sqrt{\frac{p(1-p)}{n}}\right) \leq \alpha, \tag{19}$$

where z_α is a threshold under which we find all realizations of Z, with probability $(1 - \alpha)$.

Therefore, from (19) we can assert with probability $(1 - \alpha)$ that:

$$p - \hat{p} < \varepsilon(n, \alpha), \tag{20}$$

with

$$\varepsilon(n, \alpha) = z_\alpha \sqrt{\frac{p(1-p)}{n}}. \tag{21}$$

Assume that we want to fix $\varepsilon(n, \alpha)$ to a given fraction β of p:

$$\varepsilon(n, \alpha) = \beta p. \tag{22}$$

TABLE 2
TEST SET SIZES NEEDED FOR I.I.D. ERRORS: TABLE OBTAINED BY APPROXIMATING THE BINOMIAL LAW WITH THE NORMAL LAW

p	0.01			0.03			0.1		
$\beta\backslash\alpha$	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
0.1	53,746	26,952	16,220	17,553	8,803	5,297	4,886	2,450	1,474
0.2	13,436	6,738	4,055	4,388	2,201	1,324	1,221	612	368

We assume that the best recognizer will not have an error rate p *lower than 1 percent, 3 percent, or 10 percent. With such test set sizes, with risk* α *of being wrong, the expected value of the error rate will not be worse than* $1/(1-\beta)$ *times the empirical test error rate* $\hat{p}$.

From (20), (21), and (22), we can assert, with risk α of being wrong that a number of examples:

$$n = \left(\frac{z_\alpha}{\beta}\right)^2 \frac{(1-p)}{p} \tag{23}$$

is sufficient to guarantee that the expected value of the error rate p is not worse than $\hat{p}/(1-\beta)$. To use this formula, p needs to be estimated from the results of previous benchmarks and z_α can be taken from Table 1 or conveniently approximated by $z_\alpha \simeq \sqrt{-\ln \alpha}$.

For small values of p, we will use the simplified formula:

$$n = \left(\frac{z_\alpha}{\beta}\right)^2 \frac{1}{p}. \tag{24}$$

The validity of the approximation of the binomial law by the Normal law in the tail of the distribution is questionable, even for large values of the product np. However, a bound due to Chernoff [3] asserts that with probability $(1-\alpha)$:

$$p - \hat{p} < \sqrt{-2\ln\alpha}\sqrt{\frac{p}{n}}. \tag{25}$$

Following a similar derivation as above, the number of examples needed to satisfy this more pessimistic bound is:

$$n = \frac{-2\ln\alpha}{\beta^2 p}. \tag{26}$$

By comparing (24) and (26), we see that, at worst, the approximation of the binomial law by the Normal law suggests the use of a test set which is two times too small.

For practical purposes, we will use a simplified formula, which lies between the Normal law and the pessimistic bound, obtained for typical values of α and β ($\alpha = 0.05$ and $\beta = 0.2$):

$$n \simeq \frac{100}{p}. \tag{27}$$

3.1.1 Numerical Application

From (27), for small values of p, n is inversely proportional to p. Therefore, the choice of n (the number of test samples needed) is determined by the smallest error rate which is provided by the best recognizer.

A survey of the handwriting recognition literature and of the results of recent benchmarks [5], [6], [7] indicates that the best recognizers of isolated handwritten characters will probably not have a character error rate lower than 1 percent ($p = 0.01$).

For $p = 0.01$, we obtain:

$$n \simeq 10{,}000 \text{ characters} \tag{28}$$

Performance of word recognizers using lexicons vary a lot depending on the size of the lexicon. For a task of intermediate difficulty, such as the recognition of handprinted characters with a 25,000 word vocabulary, the best recognizers will probably not have a word error rate lower than 3 percent ($p = 0.03$).

For $p = 0.03$, we obtain:

$$n \simeq 3{,}000 \text{ words} \tag{29}$$

It is important to note that the above derivation and results do not depend on the number of classes being recognized. In fact, to get statistically meaningful results, it may not even be necessary to have samples of all the classes in the test. For example, in the word recognition example given above, the number of test words that is recommended is only 3,000 words, even if the vocabulary is 25,000 words. However, the 3,000 words must be obtained randomly from a variety of writers.

The suggested test sizes given above assume that the data/errors are i.i.d., which they are not in practice. In a realistic test, where the data/errors are correlated, the required number of test examples increases somewhat, as we will see in the next section.

4 TEST SET SIZE NEEDED WHEN THE WRITER DIVERSITY IS LIMITED

The variability of the test results is affected by a number of parameters, including number of writers, conditions of data collection, and choices of test material. The theoretical solution to that problem when designing a test set is to vary as much as possible these parameters to reflect "all" the situations which could arise in the "real" world. In practice, we have little or no handle over most parameters. The solution which was adopted for the UNIPEN project [2] is to gather data collected by a large number of institutions and therefore obtain a variety of writers, conditions of data collection, and choices of test material. There is enough data that we can consider splitting it into several test sets and a training set. Our strategy is to use data from every institution to maximize the variety of conditions of data collection and choices of test material. The problem reduces to finding how many writers and how many examples per writer should go into each set, knowing that data are valuable for training and that we want to keep the test sets as small as possible.

In this section, we assume that the data are drawn from a double random process: first a writer i is picked at random from an unknown probability distribution $P(writer_i)$. Then, an example is drawn at random according to another unknown

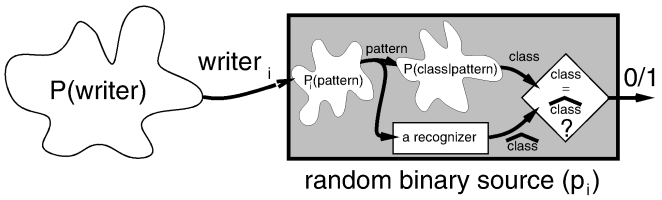

Fig. 3. Recognition of multiwriter data: We consider a double random process. A writer *i* is first picked at random. Then a pattern is picked from that writer's distribution $P_i(pattern)$.

probability distribution $P_i(pattern)$. $P_i(error) = P(error \mid writer_i)$ is then fully determined from $P_i(pattern)$, $P(class \mid pattern)$, and the recognizer (Fig. 3). The overall error distribution is given by $P(error) = \sum_i P(writer_i)P_i(error)$.[3] We will assume that $P_i(error)$ still follows a Bernoulli process, with probability p_i that the recognizer makes an error and $(1 - p_i)$ that it recognizes correctly. We will assume that $P_i(error)$ providing the probability of error p_i is distributed according to the Normal law of mean p and variance σ.[4]

The direct solution to this problem would be to compute guaranteed estimators of the error rate for the distribution $P(error) = \sum_i P(writer_i)P_i(error)$. We simplify the problem by calculating first the number of writers needed to guarantee a good estimate of the mean, neglecting the uncertainty on the writer means. We then estimate the minimum number of examples per writer needed. The considerations developed in this section have strong connections with the analysis of variance (ANOVA) statistical test [1].

4.1 Number of Writers

We call X_{ij} the random variables the realizations of which are the indicators x_{ij} (x_{ij} = 0 or 1) of the errors made by a given recognizer on examples *j* obtained from writers *i*.

We introduce further the notation $X_{i\cdot} = (1/n_w)\sum_{j=1}^{n_w} X_{ij}$ for the writer mean over n_w examples.[5] The expected value of $X_{i\cdot}$ is p_i and its variance, the "within-writer" variance, is $p_i(1 - p_i)/n_w \simeq p_i/n_w$. Realizations of $X_{i\cdot}$ are called $\hat{p}_i$.

We denote by $X_{\cdot\cdot} = (1/m)\sum_{i=1}^{m} X_{i\cdot}$ the global mean over m writers. The expected value of $X_{\cdot\cdot}$ is called p and a realization of it is $\hat{p}$.

We call σ^2 the variance of the $X_{i\cdot}$s, also called the "between-writer" variance. It is the expected value of $(X_{i\cdot} - p)^2$ over writers drawn according to their underlying distributions, for test sets of n_w examples per writer, also drawn according to their underlying distribution. An estimate of this quantity is given by:[6]

$$\hat{\sigma}^2 \simeq \frac{\sum_{i=1}^{m} (\hat{p}_i - \hat{p})^2}{m}. \tag{30}$$

It is important for the discussion that will follow to notice that σ is not the expected value of $(p_i - p)^2$. When all writers are identical ($p_i \equiv p$), this last quantity is null, whereas what we call the "between-writer" variance σ^2 tends to $p(1 - p)/n_w$ (the variance of the mean error of a given writer, that we call "within-writer" variance).

Since $X_{\cdot\cdot}$ is the mean over m writers of $X_{i\cdot}$, its variance is σ^2/m. Under the assumption that the writer error rates are Normally distributed, the random variable:

$$Z = \frac{p - X_{\cdot\cdot}}{\sigma/\sqrt{m}} \tag{31}$$

obeys the standardized Normal law (with mean 0 and variance 1).

With a risk α of being wrong, we have:

$$p - \hat{p} < z_\alpha \frac{\sigma}{\sqrt{m}} \tag{32}$$

where $\hat{p}$ is a realization of $X_{\cdot\cdot}$, and z_α a threshold obtained from table of the Normal distribution (see Table 1).

This provides us with a *guaranteed estimator* of the average error rate per writer:

$$p - \hat{p} < \varepsilon(m, \alpha) \tag{33}$$

with:

$$\varepsilon(m, \alpha) = z_\alpha \frac{\sigma}{\sqrt{m}}. \tag{34}$$

Assume that we want to fix $\varepsilon(m, \alpha)$ to a given fraction of p:

$$\varepsilon(m, \alpha) = \beta p. \tag{35}$$

From (34) and (35), we can assert, with risk α of being wrong, that a number of writers:

$$m = \left(\frac{z_\alpha \sigma}{\beta p}\right)^2 \tag{36}$$

is sufficient to guarantee that the expected value of the average error rate across writers is not worse than $\hat{p}/(1 - \beta)$, where p is the expected error rate, σ^2 is the "between-writer" variance and z_α can be taken from Table 1 for given risks α of underestimating m. As before, z_α can be conveniently approximated by $z_\alpha \simeq \sqrt{-\ln \alpha}$.

4.1.1 Numerical Application

We remind the reader that σ is a function of the number of examples per writer n_w. However, for large values of n_w, it is largely independent of n_w.

In Table 3, we calculated estimates of the number of writers needed for various values of α and β and the ratio σ/p. In Fig. 4, we show a plot of $\hat{\sigma}$ versus $\hat{p}$ for data obtained from the NIST benchmark of OCR for isolated handwritten characters [5]. The "between-writer" standard deviation of those data $\hat{\sigma}$ lies roughly between 0.5 $\hat{p}$ and $\hat{p}$. In [8], the authors

3. We sometimes talk about "correlations" between errors, hinting that errors might depend on one another. We can either view the problem as an i.i.d. problem with a more complex overall distribution or as a non-i.i.d. problem for which drawing a pattern from a particular writer increases the chance of drawing again a pattern from the same writer. For simplicity, we treat the problem as an i.i.d. problem with a more complex overall distribution.

4. This is a rather strong assumption. In general, nothing allows us to assert that this is true. Since the distribution of average writer error rates in unknown, it is difficult to find good *guaranteed estimators* which do not make simplifying assumptions. The empirical distributions provided in [5] indicate that in the case of isolated handwritten character recognition, this assumption is more or less reasonable.

5. For simplicity, we assume that all the writers have the same number n_w of examples per writer.

6. We neglect the corrective terms that arise because the means are estimated from data (see [1] for details).

TABLE 3
NUMBER OF WRITERS NEEDED

σ/p	0.5			1			2		
$\beta\backslash\alpha$	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
0.1	136	68	41	543	272	164	2,172	1,089	655
0.2	34	17	10	136	68	41	543	272	164

It is assumed that the best recognizer will have an expected character error rate p *and a "between-writer" standard deviation around* $\sigma \simeq 0.5$p, $\sigma \simeq$ p, *or* $\sigma \simeq 2$p. *Using the prescribed number of writers, with risk* α *of being wrong, the expected value of the error rate will not be worse than* $\hat{p}/(1-\beta)$.

also report a "between-writer" standard deviation which is of the order of the mean. Therefore, we adopted the value:

$$\sigma \simeq p \tag{37}$$

in our calculations.

We know that $\sqrt{p/n_w}$ is a lower bound of σ. Therefore, with the value $\sigma \simeq p$, our hypothesis that σ is largely independent of n_w will be verified when $n_w \gg 1/p$.

It is unclear whether the ratio σ/p is affected by changes in the classes of interest (e.g., words instead of characters) and whether this result applies to speech as well. We hope that new benchmark results will allow us to refine that value in the future.

Assuming $\sigma \simeq p$, we obtain the simplified formula:

$$m = \left(\frac{z_\alpha}{\beta}\right)^2 \tag{38}$$

With $\sigma = p$, for 68 writers, with 95 percent confidence ($\alpha = 0.05$), the expected value of the error rate is not worse than 1.25 times the error rate of the best recognizer ($\beta = 0.2$). One needs to double the number of writers to get 99 percent confidence ($\alpha = 0.01$) and to multiply it by four to decrease the margin of error to 0.1 ($\beta = 0.1$). In the following, we adopt:

$$m \simeq 100 \text{ writers} \tag{39}$$

4.2 Number of Examples Per Writer

We examine now the problem of determining the number of examples per writer. In the numerical examples, we fix the values of α and β to $\alpha = 0.05$ and $\beta = 0.2$, the number of writers to $m \simeq 100$ and the error rate of the best recognizer to $p = 0.01$. In each subsection below, we make a different assumption and derive a different requirement.

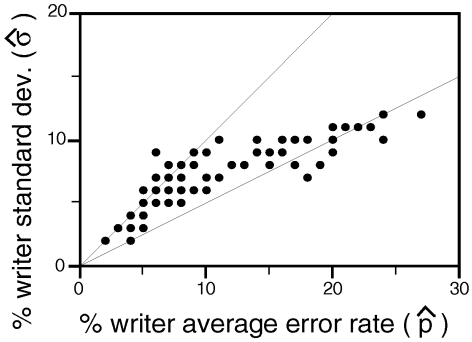

Fig. 4. Between writer variance as a function of the error rate: Each point represents the results of one recognizer from the benchmark of isolated handwritten characters published by NIST in 1992 [5]. A strong correlation between the between-writer variance and the error rate is observed.

4.2.1 Each Writer Error Rate Is Statistically Significant

The most stringent criterion is to ask for the error rate for each writer to be individually statistically significant. For instance, if we use $\alpha = 0.05$, $\beta = 0.2$, and $p = 0.01$, we obtain from (27) a number of characters per writer of approximately:

$$n_w \simeq 10{,}000 \text{ characters / writer} \tag{40}$$

and the total number of characters comes to $n' = mn_w = 100 \times 10{,}000 = 1{,}000{,}000$. With this calculation,

$$n' \gg n \tag{41}$$

where n is calculated according to (27), using i.i.d. hypotheses ($n = 10{,}000$). Note that, unless the goal is to estimate individual writer error rates accurately, n' is an overestimate of the number of characters needed.

4.2.2 All Writers Are Identical

The other extreme is to ignore the correlations between examples of the same writer and make the assumption that all writers are identical. This means that the expected value of the error rate of a given recognizer has the same value p for all writers. Let us further assume that the test set is composed of identical size subsets of n_w examples per writer. We call $\hat{p}_i$ the empirical error rate of a given recognizer for writer i. The differences between $\hat{p}_1, \hat{p}_2, \ldots$ are only due to the fact that they are estimated on a limited data set of size n. These differences are reflected by the "within-writer" variance $p(1-p)/n_w$. It is known from the ANOVA test [1] that when all the writers are identical, the expected values of the "within-writer" variance and the "between-writer" variance are equal. By replacing $\sigma^2 = p(1-p)/n_w$ in (36), we obtain $m = n / n_w$, where n is given by (24). Consequently, the total number of examples is the same as the one calculated for i.i.d. errors:

$$n' = n \tag{42}$$

For $\alpha = 0.05$, $\beta = 0.2$, and $p = 0.01$, the number of characters of 10,000 obtained from (27) is the *total* size of the test set $n' = n = mn_w$. The number of characters per writer is only $n_w = n/m = 10{,}000 \div 100$:

$$n_w = 100 \text{ characters / writer} \tag{43}$$

4.2.3 Balance Between "Within-Writer" and "Between-Writer" Variance

We make now the more realistic assumption that the empirical writer error rates are random variables X_i., normally distributed with mean p and variance σ^2. In our notation, p

TABLE 4
NUMBER OF EXAMPLES PER WRITER

γ	100	50	20	10	5	2	1
n_w	10,000	5,000	2,000	1,000	500	200	100

It is assumed that the best recognizer will not have an average character error rate p *lower than* 1 *percent and will have a "between-writer" standard deviation also around* 1 *percent.*
γ is the ratio between the "between-writer" variance and the "within-writer" variance.

is the expected value of the writer error rates $\hat{p}_i$ (which are no longer identical), σ^2 is the "between-writer" variance, an estimate of which is given by (30).

The number of examples per writer n_w can be expressed as a function of the ratio γ of the "between-writer" variance σ^2 and the "within-writer" variance $p(1-p)/n_w$. For small error rates, the "within-writer" variance can be approximated by p/n_w. We define a new parameter:

$$\gamma = \frac{n_w\sigma^2}{p(1-p)} \simeq \frac{n_w\sigma^2}{p} \tag{44}$$

The number of examples per writer as a function of γ is given by:

$$n_w = \frac{\gamma p}{\sigma^2} \tag{45}$$

From (36) and (45), the total number of examples given by $n' = mn_w$ is:

$$n' = \gamma n \tag{46}$$

where n is the number of examples calculated for i.i.d. errors (24). Notice that the case $\gamma = 1$ corresponds to having all writers identical. Testing whether γ is significantly different from 1 is the basis of the ANOVA test [1].

4.2.4 Numerical Application

In Table 4, we give the values of the number of examples per writer n when γ varies, for $p = \sigma = 0.01$.

- For $\gamma = 1$, we find again the number of characters per writer which assumes all writers are identical (see Section 4.2.2).
- For $\gamma = 100$, we get approximately the number of characters per writer which ensures that, with risk $\alpha = 0.05$, the error rate of each writer p_i is no more than $1.25\,\hat{p}_i$, which corresponds to $\beta = 0.2$. (see Section 4.2.1).
- Experts in character recognition suggest to take a number of characters per writer around:

$$n_w = 1{,}000 \text{ characters / writer} \tag{47}$$

which corresponds to $\gamma = 10$. For 100 writers, a test set size of $n' = mn_w = 100{,}000$ characters would be obtained. Note that, with such a value of γ and with risk $\alpha = 0.05$, the error rate p_i of each writer individually has a larger error bar ($\beta = 0.5$).

In practice, the number of examples per writer n_w may be given. In this case, the number n of examples using the i.i.d. assumption is first determined. From n_w and estimates of p and σ, γ is calculated. The total number of examples is then calculated from (46). From the experimental data shown in Fig. 4, if σ is unknown, γ can be approximated by:

$$\gamma \simeq n_w p. \tag{48}$$

Since γ cannot be smaller than one, by definition, we will use:

$$\gamma \simeq \max(1, n_w p) \tag{49}$$

4.3 Generalization to Multiple Factors of Correlation Between Errors

Variations in writers is only one of many possible factors of error correlation. Other factors φ may include variations in recording conditions, variations in linguistic material, etc. (see Fig. 5). Various coefficients γ_φ may be calculated to take these various factors into account. From (33) and (34), and using the approximation $z_\alpha = \sqrt{-\ln \alpha}$, for each correlation factor taken separately, one should satisfy:

$$\text{Prob}\left(p - \hat{p} \geq \sqrt{-\ln \alpha}\, \frac{\sigma}{\sqrt{m_\varphi}} \right) \leq \alpha . \tag{50}$$

If n' is the total number of examples and m_φ is the number of values taken by the correlation factor considered (e.g., φ = writer, m_φ = number of writers), then $n_\varphi = n'/m_\varphi$ is the number of examples for each value (e.g., n_φ = number of examples per writer).

By introducing $\gamma_\varphi = n_\varphi \sigma^2 / p$, we obtain:

$$\text{Prob}\left(\frac{p - \hat{p}}{\sqrt{p}} \geq \sqrt{-\ln \alpha}\, \sqrt{\frac{\gamma_\varphi}{n'}} \right) \leq \alpha . \tag{51}$$

Let us call N_φ the total number of factors. In principle, a problem of N_φ factors of correlation between errors is a N_φ dimensional problem. We first assume that all the factors are independent. We further simplify the problem to N_φ one-dimensional problems and require that the conditions (51) to be satisfied simultaneously for all factors. Let us call γ_φ^{max} the largest value of γ_φ for the factors considered,

$$\text{Prob}\left(\frac{p - \hat{p}}{\sqrt{p}} \geq \sqrt{-\ln \alpha}\, \sqrt{\frac{\gamma_\varphi^{max}}{n'}} \right) \leq$$
$$\sum_\varphi \text{Prob}\left(\frac{p - \hat{p}}{\sqrt{p}} \geq \sqrt{-\ln \alpha}\, \sqrt{\frac{\gamma_\varphi}{n'}} \right) \leq N_\varphi \alpha \tag{52}$$

and, therefore, substituting $\alpha' = N_\varphi\, \alpha$ yields:

$$\text{Prob}\left(\frac{p - \hat{p}}{\sqrt{p}} \geq \sqrt{-\ln(\alpha'/N_\varphi)}\, \sqrt{\frac{\gamma_\varphi^{max}}{n'}} \right) \leq \alpha' . \tag{53}$$

We obtain the total number of examples n' satisfying a given relative error bar $\beta = (p - \hat{p})/p$, with risk α of being wrong, by solving:

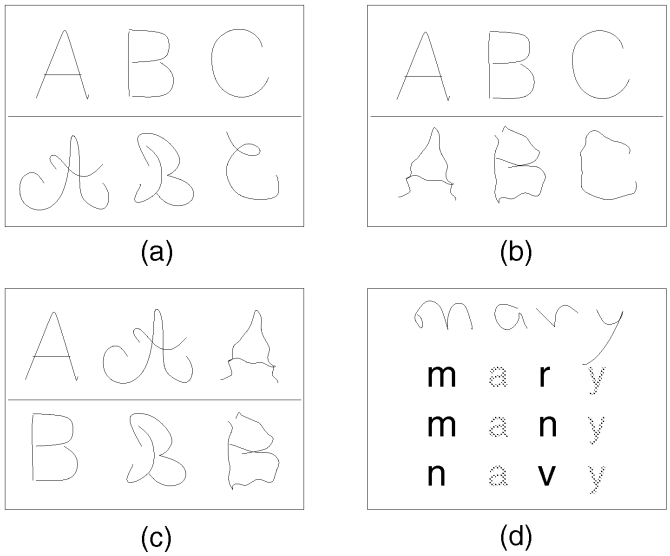

Fig. 5. Factors affecting error correlations: Errors may be correlated because of a number of factors, including (a) writing style, (b) recording conditions, (c) shape similarities within a given category, and (d) use of a language model.

$$\beta\sqrt{p} \geq \sqrt{-\ln(\alpha/N_\varphi)}\sqrt{\frac{\gamma_\varphi^{max}}{n'}} \tag{54}$$

From the value of n given by (24) with $z_\alpha = \sqrt{-\ln\alpha}$ and noticing that, for small values of α ($\alpha < 0.3$), $-\ln(\alpha/N_\varphi) \leq -(1 + \ln N_\varphi)\ln\alpha$, we obtain that with risk α of being wrong,

$$n' = \gamma_\varphi^{max}(1 + \ln N_\varphi)n \tag{55}$$

training examples guarantee that the expected value of the error rate p is not worse than $p/(1-\beta)$. Consequently, having multiple factors of error correlation increases the number of examples only with the logarithm of the number of factors, according to (55). We do not have at this point experimental data that allows us to justify our simplifying assumptions and validate this formula.

5 Test Set Sizes Which Allow Comparing the Performance of Two Recognizers

In this section, we address the problem of determining which test set size ensures that a given difference between the error rates of two recognizers is statistically significant.

We first revert to the assumption that errors are i.i.d. The method used is very simple. Since the number of common errors of the two recognizers is not known before testing, we cannot use more sophisticated methods such as the McNemar's test [9] or the method proposed in [10] which account for correlated errors. We will, however, introduce these methods in Section 7 to do a posteriori hypothesis testing.

We then address the problem of correlation between errors which is treated in a similar way as in the previous section.

5.1 The Case of i.i.d. Errors

Using similar notation as in previous sections, we call X_1 and X_2 the random variables indicating failure or success of recognition for recognizer 1 or 2 on randomly drawn examples. We call $\hat{p}_1$ and $\hat{p}_2$ their empirical error rates calculated on a test set of size n, and p_1 and p_2 the expected values of the error rates. We assume that the number of errors of both recognizers are distributed according to the binomial law, which we approximate by the Normal law. The variances of X_1 and X_2 are $var(X_i) = p_i(1-p_i)/n$, $i \in \{1, 2\}$.

Our goal is to find the smallest number of test examples n needed to assert, with a certain confidence, that recognizer 1 is better than recognizer 2, for a given difference in their error rates $\hat{p}_2 - \hat{p}_1 > 0$. This can be formalized as determining the smallest number of examples n such that, with risk α of being wrong, we can reject the hypothesis H_0 that $p_1 = p_2$ for a given value of $\hat{p}_2 - \hat{p}_1 > 0$. The alternative hypothesis H_1 that $p_2 > p_1$ is then accepted, with risk α of being wrong.

If the two random variable X_1 and X_2 are independent, $var(X_2 - X_1) = var(X_2) + var(X_1)$. If we further make the hypothesis H_0 that $p_1 = p_2 = p$, then

$$\mathrm{var}(X_2 - X_1) = 2\frac{p(1-p)}{n}. \tag{56}$$

For small values of p, we have:

$$\mathrm{var}(X_2 - X_1) \simeq \frac{2p}{n}. \tag{57}$$

Under our approximations, if the hypothesis H_0 is true, then the random variable

$$Z = \frac{X_2 - X_1}{\sqrt{2p/n}} \tag{58}$$

obeys the standardized Normal law (of mean 0 and variance 1). Therefore:

$$\mathrm{Prob}\left(\frac{\hat{p}_2 - \hat{p}_1}{\sqrt{2p/n}} \geq z_\alpha\right) \leq \alpha, \tag{59}$$

where z_α can be determined from tables of the Normal law (see Table 1). Thus, if

$$\hat{p}_2 - \hat{p}_1 \geq z_\alpha\sqrt{2p/n}, \tag{60}$$

we will reject H_0, with risk α of being wrong, and declare that recognizer 1 is significantly better than recognizer 2. Conversely, if we impose a given relative difference:

$$\beta = \frac{\hat{p}_2 - \hat{p}_1}{p}, \tag{61}$$

where $p = (p_1 + p_2)/2$, then to determine whether recognizer 1 is significantly better than recognizer 2, we need a minimum number of test examples of:

$$n = \left(\frac{z_\alpha}{\beta}\right)^2 \frac{2}{p}. \tag{62}$$

It is interesting to compare this formula to (24) for which β is a bound on $(p - \hat{p})/p$.

5.1.1 Numerical Application

Assuming that the best recognizer has an error rate of $\hat{p}_1$, if the second best recognizer has an error rate of $\hat{p}_1 + \beta p$ which is only slightly worse, what size test set would allow us to conclude that #1 is better than #2? In Table 5, we vary the

TABLE 5
TEST SET SIZES NEEDED TO DIFFERENTIATE TWO RECOGNIZERS

p	0.01			0.03			0.1		
$\beta\backslash\alpha$	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
0.50	4,343	2,178	1,311	1,448	726	437	434	218	131
0.30	12,064	6,050	3,641	4,021	2,017	1,214	1,206	605	364
0.10	108,578	54,450	32,768	36,193	18,150	10,923	10,858	5,445	3,277
0.05	434,312	217,800	131,072	144,771	72,600	43,691	43,431	21,780	13,107
0.03	1.2110^6	605,000	364,089	402,141	201,667	121,363	120,642	60,500	36,409
0.01	1.0910^7	5.4410^6	3.2810^6	3.6210^6	1.8210^6	1.1010^6	1.0910^6	544,500	327,680

It is assumed that the average error rate of the two recognizers considered are p = 1 percent, 3 percent, or 10 percent. The number in the table indicates the minimum size test set which guarantees the significance of the relative difference $\beta = (\hat{p}_2 - \hat{p}_1)/p$, with risk α of being wrong. i.i.d. errors are assumed and the binomial law is approximated by the Normal law.

confidence threshold α for various values of $p = (p_1 + p_2)/2$ and $\beta = (\hat{p}_2 - \hat{p}_1)/p$.

For a typical value of α ($\alpha = 0.05$), using the notation $\Delta\hat{p} = \hat{p}_2 - \hat{p}_1$, we obtain the following simplified formula:

$$n \simeq \frac{10p}{\Delta\hat{p}^2}. \tag{63}$$

Therefore, assuming i.i.d. errors and $p = 0.01$, using

$$n \simeq 10{,}000 \text{ characters} \tag{64}$$

guarantees the statistical significance of a difference of 0.3 percent character error $(\Delta\hat{p} = 0.003)$, with 95 percent confidence ($\alpha = 0.05$). This corresponds to a relative difference $\beta = 0.3$.

5.2 Correlated Errors

When errors are correlated, it is possible to proceed like in Section 4 and introduce γ factors. The corresponding hypothesis test is called a "matched-pair" test [9]. Matched-pair tests were derived for the particular case when the correlation factor is "linguistic material": the average difference in error rates between the two recognizers are calculated on each "segment" individually, where a "segment" is typically a sentence. But the test can be generalized to other correlation factors, where a "segment" can represent all the data from one writer.

5.3 Number of "Segments"

The test data is divided into m "segments" which are homogeneous with respect to a particular correlation factor (e.g., m writers or m sentences). We are seeking the minimum value of m which guarantees that, if the error rate of recognizer 2 is larger than the error rate of recognizer 1 by a certain margin, we can assert with a certain probability that recognizer 1 is better than recognizer 2. Our hypothesis H_0 that we wish to reject is again that $p_1 = p_2 = p$.

The derivation follows steps that are similar to those in Section 4. Let us call X_1^i and X_2^i the errors of recognizers 1 and 2 on the same segment i. We introduce two random variables $\overline{X}_1^i$ and $\overline{X}_2^i$ which are the averages of X_1^i and X_2^i over n examples. Realizations of these variables are empirical error rates, $\hat{p}_1^i$ and $\hat{p}_2^i$, for segment i. We also introduce the means $\overline{X}_1$ and $\overline{X}_2$ over all segments.

If the two recognizers perform equivalently (H_0 is true), they have the same expected error rate p and between-segment variance σ^2. If $\overline{X}_1$ and $\overline{X}_2$ are independent, the variance of $\overline{X}_2 - \overline{X}_1$ is $2\sigma^2/m$. Under such set of hypotheses, the random variable:

$$\frac{\overline{X}_2 - \overline{X}_1}{\sqrt{2}\sigma/\sqrt{m}} \tag{65}$$

obeys approximately the standardized Normal law.

If H_0 is true,

$$\text{Prob}\left(\frac{\hat{p}_2 - \hat{p}_1}{\sqrt{2}\sigma/\sqrt{m}} \geq z_\alpha\right) \leq \alpha \tag{66}$$

In other words, with risk α of being wrong, if:

$$\hat{p}_2 - \hat{p}_1 \geq \sqrt{2} z_\alpha \sigma/\sqrt{m} \tag{67}$$

we can reject H_0 and assert that recognizer 1 is better than recognizer 2.

Therefore, the number of segments that guarantees the statistical significance of $\Delta\hat{p} = \hat{p}_2 - \hat{p}_1$ with probability α of being wrong, is:

$$m = 2\left(\frac{z_\alpha \sigma}{\Delta\hat{p}}\right). \tag{68}$$

Introducing the parameter $\beta = \Delta\hat{p}/p$, we have:

$$m = 2\left(\frac{z_\alpha \sigma}{\beta p}\right)^2. \tag{69}$$

It is interesting to compare this formula with (36).

5.4 Number of Examples Per Segment

Let us call n_s the number of examples per segment. If H_0 is true, the within-segment variance is $p(1 - p)/n_s \simeq p/n_s$. Similarly, as in Section 3.2.3, we define a coefficient γ, ratio of the between-segment variance σ^2 over the within-segment variance:

$$\gamma = \frac{\sigma^2}{p/n_s}. \tag{70}$$

The total number of examples is then given by:

$$n' = mn_s = \gamma n \tag{71}$$

where n is given by (62).

5.5 Generalization to Multiple Factors of Correlation

Similarly, as in Section 4.3, one can generalize to the case of multiple factors or correlation. Let us call $\gamma_\varphi^{max} = \max_\varphi \gamma_\varphi$ and N_φ the total number of factors of error correlation, one has:

$$n' = \gamma_\varphi^{max}\left(1 + \ln N_\varphi\right)n. \quad (72)$$

6 Summary and Discussion

6.1 Test Set Size Determination

In Table 6, we summarize the various steps of our method.

In practice, it is relatively easy to obtain values for p, β_A, β_B, n_φ, N_φ, and α, but the values of σ_φ might be hard to guess. We can consider our method as a bootstrap method: As more results of benchmarks are available, it becomes easier to obtain a reasonable estimates of σ_φ and the calculation of the size of the test set for future benchmarks becomes more accurate. If nothing is known about σ_φ, one can assume $\sigma_\varphi \simeq p$ and use $\gamma_\varphi \simeq \max(1, n_\varphi p)$.

It is important to remember when designing a writer-independent test that if data from one writer is present in the test set, no data from that same writer should go into the training set. The same applies to other correlation factors φ when designing an φ-independent test.

In our numerical examples, we found that, if errors are i.i.d., for an error rate of $p = 0.01$ (a typical character error rate), $n = 10{,}000$ examples suffice, and for an error rate of $p = 0.03$ (a typical word error rate), $n = 3{,}000$ examples suffice. At the 95 percent confidence level ($\alpha = 0.05$), this corresponds to a relative difference $\beta_A = 0.2$. The expected value of the error rate should not exceed $1/(1 - \beta_A)\hat{p} = 1.25\hat{p}$, where $\hat{p}$ is the error rate on the test set. This also corresponds to a relative error $\beta_B = 0.3$. A difference in error rate between two recognizers of $\Delta\hat{p} = 0.3p$ is statistically significant.

To account for correlations between errors, we estimated that γ_w (φ = writer) is of the order of 10. If we assume that $\gamma_\varphi^{\max} \simeq \gamma_w$ and that $N_\varphi \simeq 4$, then, the corrected number of examples needed is: $n' = \gamma_\varphi^{\max}(1 + \ln N_\varphi)\, n \simeq 10(1 + \ln 4)\, 10{,}000 \simeq 200{,}000$.

6.2 Literature Overview

We investigated in various papers and technical reports the test set sizes that are used by pattern recognition researchers and are believed to be reasonable:

- In [5], the U.S. National Institute of Standards and Technology (NIST) organized a benchmark for Optical Character Recognition (OCR) of isolated handwritten characters. Three test sets were used, each one having 500 writers. The "digit" test set (10 classes or shape categories) had a total of 60,000 characters, the "uppercase letter" test set (26 classes) had 12,000 characters and the "lowercase letter" test set (26 classes) had also 12,000 characters. Therefore, the two letter test sets had approximately one letter / writer / class whereas the digit test set had 12 digits / writer / class. The authors mention that the first 10,000 digits were typical of the full digit test set, suggesting that this one was oversized.

These figures are related to our predictions in the following way:

- For the two letter test sets, assuming an error rate of 1 percent ($p = 0.01$), we predict that $n = 10{,}000$ characters are needed if i.i.d. errors are assumed. The i.i.d. assumption is reasonable here since the test sets contain one letter / writer / class. NIST chose test sets of $n = 12{,}000$ characters. Our prediction corroborate this choice.

TABLE 6
Summary of the Steps Taken to Determine the Test Set Size

p: Expected error rate of the best recognizer (e.g., 1 percent error gives $p = 0.01$).
φ: Factor of correlation between recognizer errors (e.g., φ = writer, φ = recording conditions, φ = linguistic constrains, φ = shape category).
σ_φ^2: Error rate variance for a given φ, when φ varies (e.g., $\sigma_{writer} \simeq p$).
n_φ: Number of examples per φ (e.g., 100 examples per writer).
N_φ: Number of factors of correlation (e.g., $N_\varphi = 4$).
α: Risk of predicting too few examples (e.g., $\alpha = 0.05$).
β_A: Guaranteed bound on the relative difference $(p - \hat{p})/p$ between the expected error rate and the empirical error rate (e.g., $\beta_A = 0.2$).
β_B: Minimum relative difference $(\hat{p}_2 - \hat{p}_1)/p$ between the empirical error rates of two recognizers to be compared that guarantees 1 is better that 2 (e.g., $\beta_B = 0.3$).
n_A: Number of test examples needed assuming i.i.d. errors for method A.
n_B: Number of test examples needed assuming i.i.d. errors for method B.
n'_A: Number of test examples needed, taking correlations into account for method A.
n'_B: Number of test examples needed, taking correlations into account for method B.
n': Total number of test examples needed, combining methods A and B.

(a)

Prepare: p, σ_φ, n_φ, N_φ, α; $\gamma_\varphi^{max} = \max_\varphi \frac{\sigma_\varphi^2}{p/n_\varphi}$ $(\gamma_\varphi \simeq \max(1, n_\varphi p))$	
Method A	Method B
Prepare: $\beta_A \geq (p - \hat{p})/p$	Prepare: $\beta_B = (\hat{p}_2 - \hat{p}_1)/p$
$n_A = \frac{-\ln\alpha}{\beta_A^2 p}$	$n_B = \frac{-2\ln\alpha}{\beta_B^2 p}$
$n'_A = \gamma_\varphi^{max}(1 + \ln N_\varphi) n_A$	$n'_B = \gamma_\varphi^{max}(1 + \ln N_\varphi) n_B$
Pick: $n' = \max(n'_A, n'_B)$	

(b)

(a) *Notations and typical values of the parameters.* (b) *Number of examples needed such that, with risk α of being wrong,* (1) *the expected value p of the error rate of the best recognizer is not worse than $1/(1 - \beta_A)$ times its empirical value $\hat{p}$ computed on the test set;* (2) *a relative difference of β_B between the error rates of two recognizers is significant.*

- For the digit test set, it is not possible to gather 10,000 i.i.d. examples because 500 writers times 10 classes make only 5,000 reasonably independent examples. Therefore we need to estimate the corrective factors. Having n_w = 120 characters per writer, we estimate $\gamma_w \simeq pn_w$ = 1.2. Only N_φ = 2 factors of correlation are involved (writer and shape category). We verify that the γ_c corresponding to the shape category correlations is smaller than γ_w. The total correlation factor is therefore: $\gamma_w (1 + \ln N_\varphi)$ = 1.2 (1 + ln2) $\simeq$ 2. Therefore, we predict that only n' = 20,000 examples would be needed for this test. The test set of 60,000 digits used by NIST is oversized, according to this prediction.
- In [6], NIST organized an OCR benchmark for Census Bureau forms. Each answer field was handprinted in uppercase letters (26 classes) and contained a few words. The test database consisted of 9,000 answer fields. Since each writer had to answer approximately 30 questions, we estimated that the database contained 300 different writers. Since each field had approximately 15 characters, we estimated that the total number of characters must have been approximately 135,000. Therefore, there was an average of 17 letters / writer / class.
- According to our prediction, assuming again a character error rate of 1 percent (p = 0.01), we predict that n = 10,000 characters are needed if i.i.d. errors are assumed. We identified N_φ = 3 obvious sources of error correlations: writer, class (or shape category) and linguistic constraints (within a given field). We estimate that the number of examples per writer is approximately n_w = 450. We have, $\gamma_w \simeq pn_w$ = 4.5. We verify that the γ_c corresponding to the shape category and the γ_l corresponding to linguistic constraints are both smaller than γ_w. Therefore, the corrective coefficient is: $\gamma_w (1 + \ln N_\varphi)$ = 4.5 (1 + ln3) $\simeq$ 9. We predict that only n' = 90,000 examples would be needed for this test. This is smaller than, but of the same order of magnitude as the NIST test set.

7 Hypothesis Testing

In this section, we summarize a number of hypothesis tests described in the literature. These tests can be used, *after the benchmark*, to verify the statistical significance of the results.

7.1 Precision of the Error Rate

In this section we assume that errors are i.i.d. and that a number n of test examples was chosen, according to (24). The best recognizer obtained an error rate $\hat{p}$ on those test examples. We first want to test the hypothesis H_0:

$$p - \hat{p} < \beta p \tag{73}$$

Equation (19) for small values of p becomes:

$$p - \hat{p} < z_\alpha \sqrt{\frac{p}{n}}. \tag{74}$$

We can further rewrite it as:

$$(p - \hat{p})^2 < \frac{z_\alpha^2}{n}(p - \hat{p}) + \frac{z_\alpha^2}{n}\hat{p}. \tag{75}$$

Solving for $p - \hat{p}$, we obtain:

$$p - \hat{p} < \frac{z_\alpha^2}{2n}\left(1 + \sqrt{1 + \frac{4n\hat{p}}{z_\alpha^2}}\right). \tag{76}$$

Therefore, if we pass the following test:

$$\frac{z_\alpha^2}{2n}\left(1 + \sqrt{1 + \frac{4n\hat{p}}{z_\alpha^2}}\right) < \beta p \tag{77}$$

we accept H_0 with risk α of being wrong. Otherwise, the number of examples n is too small to guarantee a relative error bar of β.

7.2 Comparison of Two Recognizers

In this section we assume that errors are i.i.d. and that a number n of test examples was chosen according to (62). Two recognizers have obtained error rates $\hat{p}_1$ and $\hat{p}_2$, $\hat{p}_1 < \hat{p}_2$. We first want to test the hypothesis H_0: $p_1 = p_2$. The test that we describe is analogous to the formulation of the McNemar test found in [9] and bears strong similarities with the method proposed in [10]. We present it here for clarity with notations consistent with the rest of the paper.

It is enough to compare the two recognizers on those examples where only one of the recognizers has an error. We call ν_1 and ν_2 the number of errors that each classifier makes and the other does not make. We call π_1 and $\pi_2 = 1 - \pi_1$ the *conditional* probabilities of error of each recognizer, given that *one recognizer only gives the wrong answer*.

The number ν_1 is distributed according to the Binomial law of expected value $(\nu_1 + \nu_2)\pi_1$ and variance $(\nu_1 + \nu_2)\pi_1(1 - \pi_1) = (\nu_1 + \nu_2)\pi_1\pi_2$. Similarly, ν_2 is distributed according to the binomial law of expected value $(\nu_1 + \nu_2)\pi_2$ and variance $(\nu_1 + \nu_2)\pi_2(1 - \pi_2) = (\nu_1 + \nu_2)\pi_1\pi_2$.

Let us introduce the random variable Π_2, of which $\hat{\pi}_2 = \nu_2/(\nu_1 + \nu_2)$ is a realization. Π_2 has expected value π_2 and variance $\pi_1\pi_2/(\nu_1 + \nu_2)$. For large values of ν_1 and ν_2, the random variable:

$$Z = \frac{\Pi_2 - \pi_2}{\sqrt{\pi_1\pi_2/(\nu_1 + \nu_2)}} \tag{78}$$

obeys approximately the standardized Normal law. In the particular case of $\pi_1 = \pi_2 = 1/2$ (i.e., if H_0 is true), Z becomes:

$$Z = \frac{\Pi_2 - 1/2}{\sqrt{\frac{1}{4(\nu_1 + \nu_2)}}}. \tag{79}$$

Therefore, if H_0 is true, with probability $(1 - \alpha)$, the following inequality holds:

$$\hat{\pi}_2 - \frac{1}{2} < \frac{z_\alpha}{2}\frac{1}{\sqrt{\nu_1 + \nu_2}}, \tag{80}$$

or, since $\hat{\pi}_2 = 1 - \hat{\pi}_1$:

$$\hat{\pi}_2 - \hat{\pi}_1 < \frac{z_\alpha}{\sqrt{\nu_1 + \nu_2}}. \tag{81}$$

Let us call n the total size of the test set, ν_{12} the number of common mistakes of the two recognizers, $\hat{p}_1$ the error rate of the first recognizer and $\hat{p}_2$ that of the second one. We have:

$$(\nu_1 + \nu_2)(\hat{\pi}_2 - \hat{\pi}_1) = \nu_2 - \nu_1 = \nu_2 + \nu_{12} - (\nu_1 + \nu_{12}) - n(\hat{p}_2 - \hat{p}_1). \tag{82}$$

Therefore, from (81) and (82), if

$$\hat{p}_2 - \hat{p}_1 \geq \frac{z_\alpha}{n}\sqrt{\nu_1 + \nu_2} \tag{83}$$

then, with risk α of being wrong, we can accept that recognizer 1 is better than recognizer 2.

7.3 Analysis of Variance

Finally, we may want to recalculate the coefficients γ_φ, in light of the results of the benchmark. A description of the ANOVA test of equality of the expected values can be found in [1]. We can use ANOVA to test the equality of the error rates for different values of the same factor φ (e.g., different writers). This test can be run for the various factors of correlation φ by determining whether $\hat{\gamma}_\varphi$ is significantly different from one. If $\max_\varphi \hat{\gamma}_\varphi$ is smaller than $\max_\varphi \gamma_\varphi$ that we used in our calculations of the number of test examples, our results will be known with less accuracy than we anticipated.

8 Conclusion

The number of examples in a test set should be inversely proportional to the error rate of the best recognizer. For errors independently and identically distributed (i.i.d), a rule of thumb is to use $n = 100/p$, where n is the test set size and p is the error rate of the best recognizer, as estimated, for instance, by the human error rate. This ensures that with 95 percent confidence the probability of error is not worse than 1.25 $\hat{p}$. For instance, for p = 1 percent character error rate, n = 10,000 characters are needed; for p = 3 percent word error rate, n = 3,000 words are needed.

In reality, errors are not i.i.d. because large chunks of data come from the same data collection device or from the same writer and because recognizers might make correlated errors, in particular if they use contextual information to perform recognition (e.g., language models). We examined particularly the case of correlations introduced by data coming from the same writer. If the between-writer variance is the same as the error rate p, using 100 writers ensures that with 95 percent confidence the true error rate is no more than 1.25 $\hat{p}$, where $\hat{p}$ is the empirical error rate, calculated on the test set. The number of examples per writer can be determined if the ratio γ_w of the between-writer variance to the within-writer variance is known. The size of the test set is then given by $n' = \gamma_w n$, where n is the test set size determined with the assumption that the errors are i.i.d. If N_φ factors of correlations must be taken into account, the size of the test set increases to $n' = \gamma_\varphi^{\max}(1 + \ln N_\varphi)n$ examples. Typical values are $1 \leq \gamma_\varphi^{\max} \leq 10$ and $1 \leq N_\varphi \leq 4$.

We examined the number of examples needed to be able to discriminate between two recognizers with very close error rates. To ensure that with 95 percent confidence a difference $\Delta\hat{p}$ in error rate is significant, the test set size must exceed $n = 10p/\Delta\hat{p}^2$ i.i.d. examples. For a difference $\Delta\hat{p}$ = 0.3 percent, approximately n = 10,000 examples are needed if this rule is followed. If data or errors are not i.i.d., corrections can be performed with multiplicative coefficients taking into account correlations, as explained above.

These guidelines should provide some insight for benchmark organizers. Of course, this simplified framework might not be strictly applicable in all situations. For instance, the number of examples per writer might vary substantially from writer to writer in a given database. In such a case, a specific data splitting algorithm must be derived, keeping in mind the general principle:

1) maximize data diversity in the test set, with respect to data source, shape categories and linguistic material;
2) in a writer/speaker independent task, forbid data from the same writer/speaker to be both in the training and test sets;
3) impose a minimum number of 100 writers/speakers;
4) reach the minimum number of examples prescribed.

Finally, we should emphasize that nowhere did we make the hypothesis that our test set sizes were for writer/speaker independent tasks only. They apply as well to writer/speaker dependent tasks, in which data from the same writer is both in the training and the test set.

Acknowledgments

We would like to thank our colleagues Lambert Schomaker of the NICI in the Netherlands and Stan Janet of the U.S. NIST for their helpful suggestions on how to improve this paper and Réjean Plamondon of l'Ecole Polytechnique de Montréal for pointing out several references.

Part of this work was done while Isabelle Guyon was working at AT&T Bell Laboratories, Holmdel, New Jersey.

References

[1] A.M. Mood, F.A. Graybill, and D.C. Boes, *Introduction to the Theory of Statistics*. McGraw Hill, 1974.

[2] I. Guyon, L. Shomaker, R. Plamondon, M. Liberman, and S. Janet, "UNIPEN Project of On-Line Data Exchange and Benchmarks, *Proc. 12th Int'l Conf. Pattern Recognition*, IAPR-IEEE, 1994.

[3] H. Chernoff, "A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sums of Observations," *Annals of Mathematical Statistics*, vol. 23, pp. 493-509, 1952.

[4] W. Hoeffding, "Probability Inequalities for Sums of Bounded Random Variables," *J. Am. Statistics Assoc.*, vol. 58, pp. 13-30, 1963.

[5] R. A. Wilkinson, J. Geist, S. Janet, P.J. Grother, C.J.C. Burges, R. Creccy, B. Hammond, J.J. Hull, N.J. Larsen, T.P. Vogl, and C. Wilson, "The First Census Optical Character Recognition Systems Conference," Technical Report NISTIR-4912, NIST, U.S. Dept. of Commerce, 1992.

[6] J. Geist, R.A. Wilkinson, S. Janet, P.J. Grother, B. Hammond, N.W. Larsen, R.M. Klear, M.J. Matsko, C.J.C. Burges, R. Creecy, J.J. Hull, T.P. Vogl, and C. Wilson, "The Second Census Optical Character Recognition Systems Conference," Technical Report NISTIR-5452, NIST, U.S. Dept. of Commerce, 1994.

[7] I. Guyon, M. Schenkel, and J. Denker, *Overview and Synthesis of On-Line Cursive Handwriting Recognition Techniques*. World Scientific, in press.

[8] I. Guyon, D. Henderson, P. Albrecht, Y. Le Cun, and J. Denker, "Writer Independent and Writer Adaptive Neural Network for On-Line Character Recognition," S. Impedovo, ed., *From Pixels to Features III*, pp. 493-506. Amsterdam: Elsevier, 1992.

[9] L. Gillick and S.J. Cox, "Some Statistical Issues in the Comparison of Speech Recognition Algorithms," *Proc. ICASSP*, IEEE 1989.

[10] L. Bottou and V. Vapnik, "Local Learning Algorithms," Technical Report TM-11359-920124-05, AT&T Laboratories, Holmdel, N.J., 1992.

Isabelle Guyon received an engineering diploma from the Ecole Superieure de Physique et Chimie Industrielle de Paris, in 1985 and PhD in physical sciences from the Universite Pierre et Marie Curie, Paris, in 1988. She was a research scientist at AT&T Bell Labs Research, New Jersey, from 1989 to 1996. She now has her own company, ClopiNet, which provides consulting services. Dr. Guyon's work focuses on learning systems and, in particular, neural networks and Markov models. She has also strong interests in pattern recognition algorithms in general, classical statistics, and learning theory. She has coauthored many papers in handwriting recognition.

John Makhoul is an alumnus of the American University of Beirut and the Ohio State University, and received his PhD in electrical engineering from MIT in 1970, with specialization in speech recognition. He is currently a chief scientist at BBN Systems and Technologies, Cambridge, Massachusetts. He is also an adjunct professor at Northeastern University and the University of Massachussetts, Boston, and a research affiliate at the MIT Speech Communication Laboratory. He has been with BBN directing various projects in speech recognition, spoken language systems, speech coding, speech synthesis, speech enhancement, signal processing, neural networks, and character recognition. Dr. Makhoul is a fellow of the IEEE and the Acoustical Society of America.

Richard Schwartz received an S.B. degree in electrical engineering from MIT. He joined BBN Systems and Technologies, Cambridge, Massachusetts in 1972, and is currently a principal scientist. He specializes in speech recognition, speech synthesis, speech coding, speech enhancement in noise, speaker identification and verification, neural networks, statistical language understanding, and character recognition.

Vladimir Vapnik, Technology Consultant AT&T Labs-Research, is one of the creators of generalization theory in statistical inference, the so-called VC-theory (abbreviation for the Vapnik-Chervonenkis theory). Dr. Vapnik is the author of many articles and books devoted to different problems of the statistical learning theory.