

On the editing rate of the MULTIEDIT algorithm

Pierre A. DEVIJVER

Philips Research Laboratory, Avenue Van Becelaere 2, Box 8, B-1170 Brussels, Belgium

Received 11 September 1985

Abstract: In a number of previous publications, we have stated, without proof, that the total fraction of samples discarded by the MULTIEDIT algorithm is bounded from above by $2E_1$, where E_1 is the 1-NNR error rate for the initial distributions. It is the purpose of this note to offer a more precise formulation together with a derivation of this assertion.

Key words: Edited nearest neighbor rule, MULTIEDIT algorithm, editing rate.

1. Introduction

The so-called MULTIEDIT algorithm is an efficient preprocessing technique which permits to remove from a training set all those samples that do not belong to the Bayes acceptance region of their class. In this way, it permits to achieve minimal (Bayes) error probability by 1-NN classification with a MULTIEDITED training set. It also proves useful prior to applying condensing techniques as well as in other contexts (Voisin (1985)).

A detailed analysis of the MULTIEDIT algorithm can be found in Devijver and Kittler (1980, 1982). In subsequent publications (e.g., Devijver (1982, 1984)) we have stated without proof that the total fraction of samples edited upon termination of MULTIEDIT is bounded from above by $2E_1$, where E_1 is the 1-NNR error rate for the initial distributions. It is the goal of this note to offer a more precise formulation, together with a derivation of this assertion. The property we wish to establish is stated precisely at the end of this section. The derivation is the subject matter of the next two sections. Section 4 contains an example and some comments. Presently, we briefly introduce the required definitions and notation.

We assume a problem of m pattern classes with a priori and conditional probabilities P_i and $p(X/\omega_i)$ respectively and a posteriori probabilities $\eta_i(X) =$

$P_i p(X/\omega_i)/p(X)$ where $p(X) = \sum_i P_i p(X/\omega_i)$, $i=1, \dots, m$. Cover and Hart (1967) have shown that, under the large-sample assumption, the error rate E_1 of the 1-NN decision rule is given by

$$E_1 = \int 2 \sum_{i < j} \eta_i(X) \eta_j(X) p(X) dX. \quad (1)$$

The MULTIEDIT technique permits to make repeated use of Wilson's (1972) editing idea in an independent manner. One iteration of the algorithm consists in four steps:

1. Make a random partition of the available training data into M subsets, $S(1), \dots, S(M)$.
2. Classify the sample in $S(i)$ using the 1-NNR with $S((i+1) \bmod M)$ as a training set, $i=1, \dots, M$.
3. Discard all the samples that were misclassified at Step 2.
4. Pool the remaining data to constitute a new training data set.

An asymptotic analysis of this scheme shows that its effect is to produce apparent underlying distributions from which the remaining data could have been drawn. So, let $\eta_i^{(n)}(X)$, $i=1, \dots, m$, and $p^{(n)}(X)$ designate the apparent a posteriori probabilities and unconditional density at the start of the n th iteration. It was shown in Devijver and Kittler (1982) that

$$\eta_i^{(n)}(X) = \frac{[\eta_i^{(n-1)}(X)]^2}{1 - e_1^{(n-1)}(X)} \quad (2)$$

and

$$p^{(n)}(X) = p^{(n-1)}(X) \frac{1 - e_1^{(n-1)}(X)}{1 - R^{(n-1)}}, \quad (3)$$

where $e_1^{(n-1)}(X)$ is the conditional 1-NN risk under the apparent distributions for the $(n-1)$ st iteration, i.e.,

$$e_1^{(n-1)}(X) = 2 \sum_{i < j} \eta_i^{(n-1)}(X) \eta_j^{(n-1)}(X), \quad (4)$$

and $R^{(n-1)}$ denotes the corresponding 1-NN error rate. There follows $R^{(1)} = E_1$, and by analogy with (1),

$$R^{(n)} = \int e_1^{(n)}(X) p^{(n)}(X) dX. \quad (5)$$

In an actual application of the MULTIEDIT algorithm, let $N^{(n)}$, $n \geq 0$, designate the number of samples retained *at the end* of the n th iteration, with $N^{(0)}$ denoting the number of samples in the original training data. Given $N^{(n-1)}$, the expected number $N_e^{(n)}$ of samples discarded (edited) in the course of the n th iteration is

$$N_e^{(n)} = N^{(n-1)} R^{(n)}, \quad (6)$$

hence

$$N^{(n)} = N^{(n-1)} (1 - R^{(n)}). \quad (7)$$

At this point, we can characterize the behaviour of the MULTIEDIT algorithm in an accurate fashion. Specifically, in the two-class case and under the large-sample assumption for $R^{(n)}$, it holds that

$$\sum_{n=1}^{\infty} N_e^{(n)} \leq 2N^{(0)} E_1, \quad (8)$$

where it should be noted that $N^{(0)} E_1$ is the expected number of samples discarded at the first iteration. Equation (8) holds also in the multiclass case, under an additional assumption which is made precise in Section 3.

2. The two-class case

Our first step in establishing (8) will be to relate $R^{(n)}$ to $R^{(n-1)}$. To this end, we first relate the 1-NN error rates at the first and second iterations. Readily, we have

$$R^{(2)} = \int 2\eta_1^{(2)}(X) \eta_2^{(2)}(X) p^{(2)}(X) dX, \quad (9)$$

where, by (2), (3), (4) and (5),

$$\eta_i^{(2)}(X) = \frac{\eta_i^2(X)}{1 - e_1(X)}, \quad (10)$$

and

$$p^{(2)}(X) = \frac{p(X)(1 - e_1(X))}{1 - E_1}. \quad (11)$$

Note that, for the sake of simplicity, we have dropped all superscripts for quantities involving the original distributions. By substituting for $\eta_i^{(2)}(X)$ and $p^{(2)}(X)$ from (10) and (11) into (9), we get

$$R^{(2)} = \int \frac{2\eta_1^2(X) \eta_2^2(X)}{1 - e_1(X)} \frac{p(X)}{1 - E_1} dX. \quad (12)$$

In the two-class case, it is plain that

$$1 - e_1(X) \geq e_1(X) \quad (13)$$

with equality if $e_1(X) = \frac{1}{2}$. Hence,

$$\frac{2\eta_1^2(X) \eta_2^2(X)}{1 - e_1(X)} \leq \eta_1(X) \eta_2(X) = \frac{1}{2} e_1(X) \quad (14)$$

with equality if $e_1(X) = 0$ and $e_1(X) = \frac{1}{2}$. There follows

$$\begin{aligned} R^{(2)} &\leq \frac{1}{2(1 - E_1)} \int 2\eta_1(X) \eta_2(X) p(X) dX \\ &= \frac{E_1}{2(1 - E_1)} = \frac{R^{(1)}}{2(1 - R^{(1)})}. \end{aligned} \quad (15)$$

By proceeding in the same way, one would show that for any $n \geq 2$,

$$R^{(n)} \leq \frac{R^{(n-1)}}{2(1 - R^{(n-1)})}. \quad (16)$$

This is the relation we need to establish (8).

From (6) and (16), we can write

$$N_e^{(n)} \leq N^{(n-1)} \frac{R^{(n-1)}}{2(1 - R^{(n-1)})}, \quad (17)$$

while from (7) we get

$$N^{(n-1)} = N^{(n-2)} (1 - R^{(n-1)}). \quad (18)$$

By substituting for $N^{(n-1)}$ from (18) into (17), we obtain

$$N_e^{(n)} \leq \frac{1}{2} N^{(n-2)} R^{(n-1)} = \frac{1}{2} N_e^{(n-1)}. \quad (19)$$

Equation (19) shows that $N_e^{(n)}$ is bounded from above by the terms of a convergent geometric progression with first term $N^{(0)}E_1$ and ratio $\frac{1}{2}$. This establishes (8).

3. The multiclass case

In the multiclass case, the same reasoning leads to a divergent series for the upper-bound on the expected numbers of samples discarded. Still, it can be shown that (8) remains valid under an additional hypothesis which appears naturally in the derivation, and which can be easily tested by the experimenter. To see this, let us examine the multiclass extension of (12):

$$R^{(2)} = \int \frac{2 \sum_{i < j} \eta_i^2(X) \eta_j^2(X)}{1 - e_1(X)} \frac{p(X)}{1 - E_1} dX. \quad (20)$$

By the same reasoning as above,

$$\begin{aligned} 2 \sum_{i < j} \eta_i^2(X) \eta_j^2(X) &\leq \frac{1}{2} \left[2 \sum_{i < j} \eta_i(X) \eta_j(X) \right]^2 \\ &= \frac{1}{2} e_1^2(X), \end{aligned} \quad (21)$$

where equality holds if and only if, for any given X , no more than two classes have strictly positive a posteriori probabilities simultaneously. Under the provision that this assumption holds over the entire sample space we shall presently show that (8) holds.

Using (21), (20) simplifies into

$$R^{(2)} \leq \frac{1}{2(1 - E_1)} \int \frac{e_1^2(X)}{1 - e_1(X)} p(X) dX. \quad (22)$$

Let $\mu \leq m$ designate the largest number of classes that have simultaneously strictly positive a posteriori probabilities. Note that $e_1^2(X)/(1 - e_1(X))$ is a strictly concave-upwards function of $e_1(X)$ over the interval $[0, (\mu - 1)/\mu]$, and $(\mu - 1)e_1(X)$ is the least concave-downwards function of $e_1(X)$ greater than or equal to $e_1^2(X)/(1 - e_1(X))$ over that interval, uniformly in X . Combining this with (22), we obtain

$$\begin{aligned} R^{(2)} &\leq \frac{\mu - 1}{2(1 - E_1)} \int e_1(X) p(X) dX \\ &= \frac{(\mu - 1)E_1}{2(1 - E_1)}. \end{aligned} \quad (23)$$

Now, if we were to use this relationship in the place of (15) to derive an upper-bound on the expected number of samples discarded by the MULTIEDIT algorithm, it is evident that we would get a divergent geometric progression, except for $\mu = 2$. Thus, the hypothesis that, for any given X , no more than two classes have simultaneously strictly positive a posteriori probabilities appears as a necessary condition for (8) to be valid in the multiclass-case.

It should be noted that this assumption is not uncommon in multiclass analyses of the NN rule. For instance, essentially the same assumption is advocated by Short and Fukunaga (1980) in their definition of an optimal nearest neighbor distance measure.

Eventually, let us also note that, in practical applications, it is quite simple to decide whether the assumption is satisfied for a given set of training data: All that need be done is to test whether each sample and its k nearest neighbors do not belong to more than two different classes for a suitably chosen value of k .

4. Examples and comments

Extensive simulation experiments have been made to validate the theoretical analysis presented above. Experiments with artificial 2-D data from distributions for which equality was uniformly achieved in (14) exhibited extremely good concordance between experimental results and theoretical predictions. In experiments with data drawn from distributions failing to satisfy equality in (14), $N_e^{(n)}$ displayed a marked tendency towards a geometric progression with first term $N^{(0)}E_1$ and ratio less than $\frac{1}{2}$; for instance, in Devijver and Kittler (1979), ratios of 0.25, 0.26 and 0.34 can be observed.

The bound reported here may be of more than theoretical interest. For instance, the failure of experimental results to comply with the bound may call the experimenter's attention to such problems as too low a ratio of sample size to intrinsic dimensionality. We shall briefly illustrate this point on the basis of two-class experiments with speech data. In these experiments, the data consisted in 1494 acoustic vectors representing the phonemes

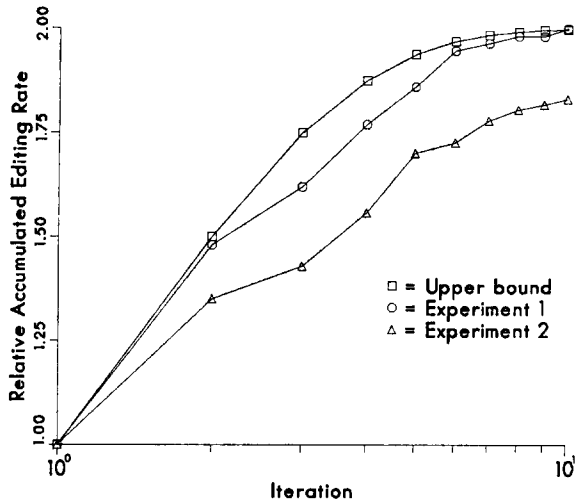


Figure 1.

“a” and “j”, e.g., the first two phonemes in the German word “eins” (one). In the first experiment, phonemes were represented by their first 16 cepstral coefficients. In the second experiment, the first 8 coefficients only were retained. Figure 1 displays a semi-logarithmic plot of the accumulated editing rates (relative to $N^{(0)}E_1$) versus iteration number, together with the bound of (8). It is seen that the curve for the first experiment follows the theoretical bound very closely, though there was no prior evidence that it ought to do so. The curve for the second experiment is a good approximation to a geometric progression with ratio 0.45, which seems more realistic. In spite of the fact that the theoretically predicted bound is not violated, these results strongly suggest that, for the problem at hand, more samples should be needed to faithfully represent the underlying distributions in \mathfrak{R}^{16} .

Eventually, it should be emphasized that the main goal of the MULTIEDIT technique is to recover the Bayes acceptance regions. The fact that this is achieved by discarding samples that do not belong to the Bayes acceptance region of their class – as well as some other samples that could legitimately

be retained – is to be regarded as purely accidental. Clearly, the larger the number of samples ultimately retained, the better the localization of the acceptance regions. In this sense, the work reported here enables us to conclude that the MULTIEDIT algorithm achieves its goal at quite a reasonable cost.

Acknowledgment

We wish to express our appreciation to Xavier Aubert who called our attention to the work of Short and Fukunaga (1980) and allowed us to incorporate some of his experimental results on speech data.

References

- [1] Cover, T.M. and P.E. Hart (1976). Nearest neighbor pattern classification, *IEEE Trans. Inform. Theory* 13, pp. 21–27.
- [2] Devijver, P.A. (1982). Advances in nonparametric techniques of statistical pattern classification. In: J. Kittler, K.S. Fu and L.F. Pau, Eds., *Pattern Recognition Theory and Applications*. Reidel, Dordrecht, pp. 3–18.
- [3] Devijver, P.A. (1984). Selection of prototypes for nearest neighbor classification. In: D. Dutta Majumder, Ed., *Advances in Information Sciences and Technology*. Indian Statistical Institute, Calcutta, pp. 84–106.
- [4] Devijver, P.A. and J. Kittler (1979). On the edited nearest neighbor rule. Philips Res. Lab. Rept. R.410.
- [5] Devijver, P.A. and J. Kittler (1980). On the edited nearest neighbor rule. *Proc. 5th Internat. Conf. Pattern Recognition*, Miami, FL, pp. 72–80.
- [6] Devijver, P.A. and J. Kittler (1982). *Pattern Recognition, A Statistical Approach*, Prentice Hall, Englewood Cliffs, NJ.
- [7] Short, R.D. and K. Fukunaga (1980). A new nearest neighbor distance measure. *Proc. 5th Internat. Conf. Pattern Recognition*, Miami, FL, pp. 81–86.
- [8] Voisin, J. (1985). Etude de la reconnaissance optique des caractères, Philips and MBLE Ass., Rept. IRSIA-137.
- [9] Wilson, D.L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Trans. Syst. Man Cybernet.* 2, pp. 408–420.