TABLE II
VALUES OF α

$(n-1)p$ \ ρ	0	0.1	0.2	0.3	0.4	0.5
4	0.99	0.988	0.986	0.983	0.979	0.971
	0.95	0.941	0.930	0.918	0.900	0.872
16	0.99	0.982	0.966	0.941	0.890	0.790
	0.95	0.919	0.870	0.790	0.670	0.490
32	0.99	0.975	0.946	0.867	0.715	0.440
	0.95	0.900	0.810	0.640	0.380	0.170

Evidently, if $\rho < 0$, $\alpha > \varepsilon$, but if $\rho > 0$, $\alpha < \varepsilon$. The effect of positive values of ρ upon the confidence coefficient α is given in Table II for $\varepsilon = 0.99$ and 0.95.

V. Conclusion

This correspondence has demonstrated that the presence of intraclass correlation affects the confidence coefficients of the confidence sets obtained under the assumption of independence for the mean of a normal population with known dispersion matrix and the dispersion scalar σ^2 in $\sigma^2\Sigma$ with Σ known. Also, the following was demonstrated.

a) If the sample is simply equicorrelated with positive coefficient of simple equicorrelation, then the confidence coefficient decreases with sample size.

b) If the sample is simply equicorrelated with negative coefficient of simple equicorrelation $(-1/(n-1) < \rho < 0)$, then the confidence coefficient increases with sample size.

Therefore, to be sure about the accuracy of the inference regarding the confidence coefficient, it is necessary to test the sample for independence or for the type of correlation—positive or negative. In case the sample is found to be positively correlated, it is advocated that test statistics appropriate for simply equicorrelated data be used. When the coefficient of positive simple equicorrelation is known or has been estimated, the test statistics appropriate for simply equicorrelated samples can be easily derived from the corresponding statistics for the independent samples.

References

[1] W. Coberly, personal communications. Memorandum from Mathematical Physics Branch of Mission Planning and Analysis Division, NASA/JSC, Houston, TX, May 1973.

[2] J. E. Walsh, "Concerning the effects of intraclass correlation on certain significance tests," *Annals of Mathematical Statistics*, vol. 18, pp. 88–96, 1947.

[3] J. P. Basu, P. L. Odell, and T. O. Lewis, "The effects of intraclass correlation on certain significance tests when sampling from multivariate normal population," *Communication in Statistics*, vol. 3, pp. 899–908, September 1974.

[4] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York: Wiley, 1958.

An Experiment with the Edited Nearest-Neighbor Rule

IVAN TOMEK

***Abstract*—A number of computer simulation experiments with the nearest-neighbor classification rule are described. They include classification by the usual *k*-NN rule, classification with *k*-NN on a design set edited once according to Wilson and classification with *k*-NN on a design set edited unlimited number of times by two methods described in the text. Results of experiments indicate that editing improves performance of the rule. This is not proved rigorously, but a possible approach to a proof is outlined.**

Introduction

Let D be a set of samples from a d-dimensional Euclidian space whose members are selected as follows. Select $m = 1$ with probability p_1 and $m = 2$ with probability $p_2 = 1 - p_1$. Given m, select $x \in D$ from a population with density $q_m(x)$. In other words, D consists of samples from two classes whose underlying density functions are $q_m(x)$ and *a priori* probabilities p_m $(m = 1,2)$. D will be called the *design set* and its members *labeled prototypes*.

The k-nearest-neighbor (k-NN) rule is a method of classification which works as follows. Given a sample y to be assigned to one of the two classes, find the k nearest neighbors (kNN) of y in D using a chosen definition of distance. Assign y to class m (either 1 or 2) if the majority of its kNN belong to class m; break ties arbitrarily. (In practice, k is usually odd to avoid ties. This will also be assumed for simplicity in the following text unless specified otherwise.)

Much research has been devoted to the k-NN rule (for a list of references see, for example, Duda and Hart [1]). One of the most important results is that k-NN has asymptotically very good performance. Loosely speaking, for a very large design set, the expected probability of incorrect classification (error) P achievable with k-NN is bounded as follows:

$$P^* \le P < 2P^*. \tag{1}$$

Here P^* is the optimal (minimal) error rate for the given underlying distributions p_i, $q_i(x)$ $(i = 1,2)$ (see [2]). In many situations the rule performs almost as well as the optimal classifier [3]. Furthermore, it has been shown by Wilson [3] that the following simple editing of D improves the performance of k-NN even further:

1) classify each sample $x(i) \in D$ by k-NN using samples $x \in D$, $x \neq x(i)$;
2) form a new design set D' containing exactly those samples from D which have been classified in accordance with their actual membership in Step 1).

k-NN classification with D replaced by D' reduces the expected error rate below that associated with D.

It is natural to ask what would similar editing of D' (leading to D''), D'' etc. do to the design set. Should we expect progressively better and better classification, or will editing distort the design set and result in deteriorating performance?

A satisfactory answer to this question has not been found yet. The reason is, basically, the following difficulty. Wilson has been able to show that if a design set is "very large" and consists of independently chosen samples, then, under quite general conditions, the asymptotic probability that a sample x classified by its k nearest neighbors in D' is assigned to class $C(1)$ is

$$\bar{P}_\infty(C(1)/x) = \frac{p_1 P_\infty(C(1)/x)}{p_1 P_\infty(C(1)/x) + p_2 P_\infty(C(2)/x)}. \tag{2}$$

Here $P_\infty(C(i)/x)$ is the asymptotic probability that sample x is assigned to class $C(i)$ by the k-NN rule with an "infinitely" large design set D.

Unfortunately Wilson's proof depends on the assumption that samples in D are chosen independently. Editing generates a design set in which samples are *not* independent since retention

Manuscript received September 12, 1975; revised January 19, 1976.
The author is with the Department of Computer Science, Acadia University, Wolfville, N.S., Canada BOP 1XO.

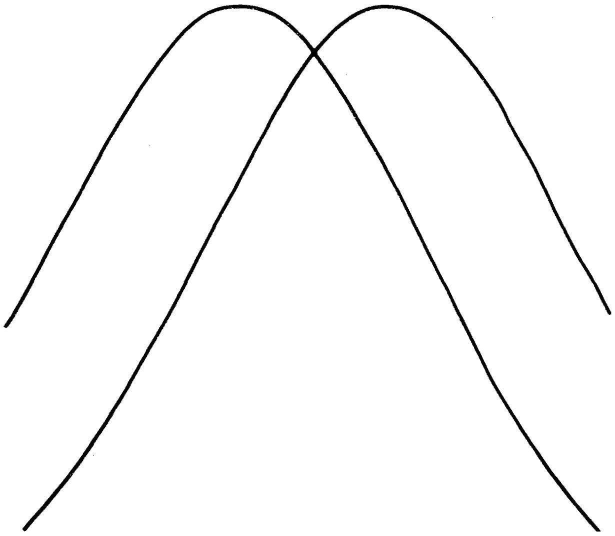


Fig. 1. Distributions used in experiment.

or elimination of samples depends upon class membership of the surrounding samples, and samples in D' are thus chosen in function of their neighbors in D'. Wilson's proof thus cannot be repeated beyond the first editing. If it could be proved that Wilson's conclusion remains valid by editing—at least for a restricted class of probability distributions—then it would be possible to show that repeated editing of a "very large" design set could in fact lead to the generation of a new design set D_∞ allowing classification as close to optimal as desired. This is shown in Appendix I.

The author has not been able to prove anything about the extension of Wilson's result to repeated editing and has thus resorted to simulation to obtain at least an indication whether this quite plausible-looking assumption should be investigated further. Results of these limited experiments (presented in the next section) are quite encouraging, particularly for a modified method of editing. This is in agreement with reasoning presented in Appendix II.

The author would like to repeat that complete theoretical justification of the main theorem presented in Appendix I is not given. Results of experiments indicate that the described methods indeed improve the performance of k-NN classification considerably, and the problem is worth further investigation. It is hoped that this paper will stimulate such research.

The Experiment

Data and Output

In order to make it possible to examine a sufficiently large number of situations in reasonable time, it was decided to examine samples taken from one-dimensional pseudonormal distributions. All experiments used distributions with means 0.0, 4.0, and standard deviations 5.0 (Fig. 1). Experiments were run with 50, 100, 200, 500, 1000, and 2000 samples in each class.

Each experiment ran as follows.

1) The desired number of samples from both classes was generated using a random number generator.
2) The generated design set was edited in several ways (described below) thus giving a number of new design sets.
3) Asymptotical performance (expected error of classification) was calculated on each set by first defining decision boundaries inherent in the design set.

Steps 1), 2), and 3) were repeated a number of times (40 for 50, 100, and 200 samples, 30 for 500 samples, 20 for 1000 samples and 10 for 2000 samples). In general, for design sets of smaller size, the number of repetitions was larger. This was dictated by the rapidly increasing processing times and the assumption that larger design sets are more representative of the underlying distributions and thus do not require that many repetitions.

Methods of Processing

The following methods of processing (editing) of the design set were used:

1) Wilson's editing (as described in the Introduction),
2) unlimited repetition of Wilson's editing (in fact, editing is always stopped after a finite number of steps because after a certain number of repetitions the design set becomes immune to further elimination),
3) editing by the "all k-NN" method. For a given value of k and a given sample x this method works essentially as follows:

```
a) $i = 1$, flag $(x) = 1$,
b) find $i$ nearest neighbors of $x$: NN $(i,x)$,
c) if the majority of NN $(i,x)$ classify $x$ incorrectly,
      flag $(x) = 0$, end.
d) $i = i + 1$,
e) if $i \leq k$ go to Step b), otherwise end.
```

After processing all samples from D, eliminate those with flag $(x) = 0$. For all combinations editing was performed for $k = 1,3,5$. Classification was performed for the same values of k.

Results and Conclusions

The most important result of the described experiments—the averaged expected error—is shown in Figs. 2, 3, and 4. Performance of various methods changes in a somewhat unexpected way—getting worse as the number of samples in the design set increases. Let us note that the performance of individual tested methods varied quite widely from one design set to another, as could be expected with the chosen type of distributions. It may be that a larger set of experiments should be performed to give a more stable behavior. This was, however, beyond the author's means.

Certain significant conclusions can be drawn from the presented results.

a) Performance improves quite significantly with increasing value of k. This is natural in view of the known results about k-NN.

b) On the average, better classification is obtained in this order:

1) unedited design set,
2) Wilson's editing,
3) unlimited editing,
4) "all k-NN."

c) The "all k-NN" method seems to be clearly superior to all considered methods. This could again be expected in view of the result about the speed of convergence shown in Appendix II.

It is hypothesized in Appendix I that k-NN editing progressively modifies the original underlying probability distributions and that these, in the limit, become disjoint in a way optimal for classification: the ratio of the likelihood of finding the nearest neighbor from a nonoptimal class to that of the nearest neighbor belonging to the optimal class approaches to zero with editing. This holds for all values of k.

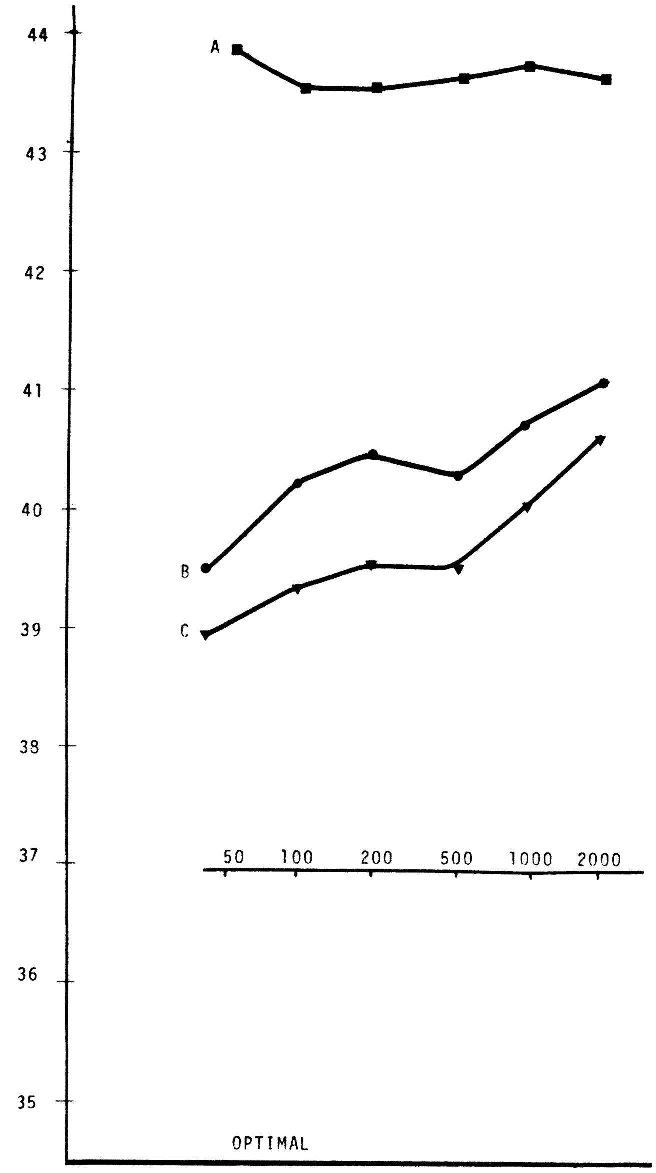


Fig. 2. Average errors of classification (%) versus number of samples in original design set. $k = 1$ A for unedited design sets, B for Wilson's editing, C for unlimited number of editing.

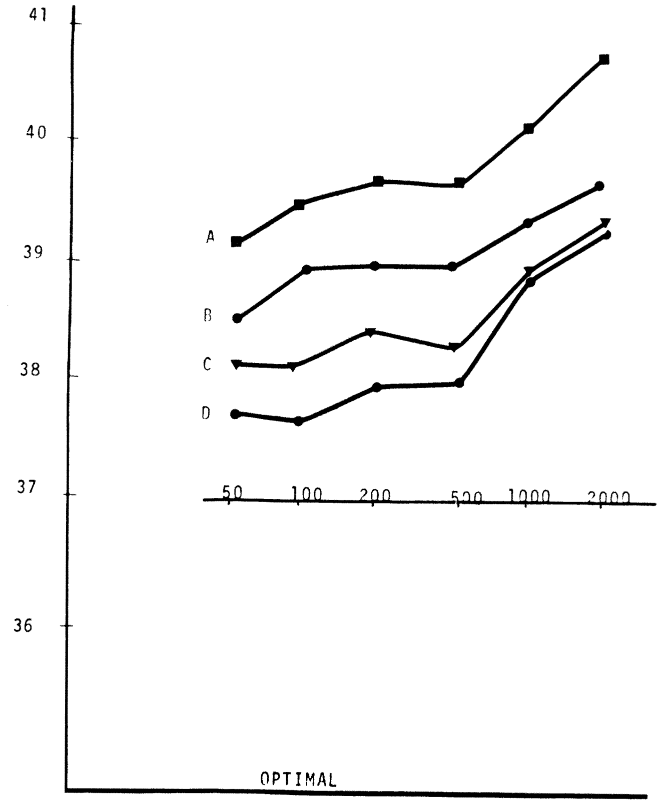


Fig. 3. Same as Fig. 2 for $k = 3$. D denotes "all NN" editing.

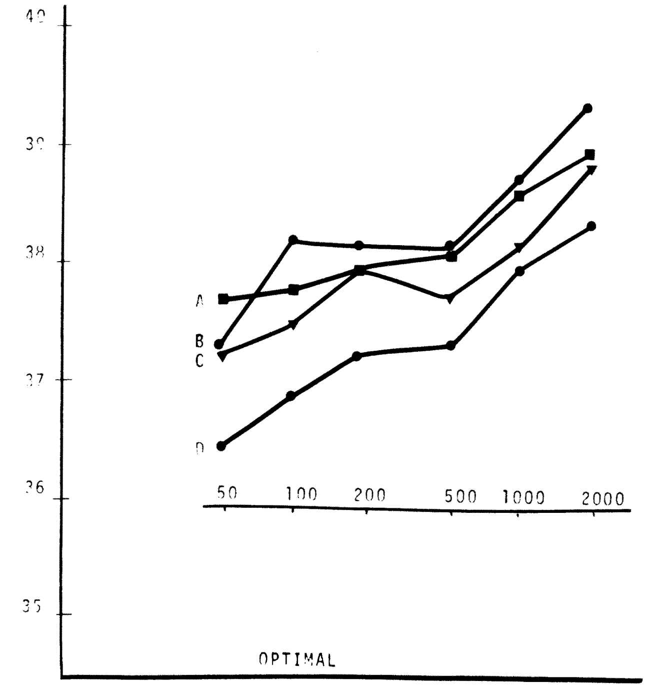


Fig. 4. Same as Fig. 3 for $k = 5$.

If, for a particular point, this ratio is denoted as α_N (for the Nth iteration) then it is shown in Appendix II that $\alpha_{N+1} \leq \alpha_N^2$ (exponential speed of "convergence," irrespective of the value of k).

For the "all NN" elimination this inequality is replaced by

$$\alpha_{N+1} \leq \alpha_N^{2^{p(k)}}$$

where

$$p(k) = \left[\frac{k+1}{2}\right]$$

(square brackets denote the integer part). The convergence, according to this formula, should be much faster than that for the ordinary elimination, and this is confirmed in our experiments.

Conclusion

Results presented in the main text indicate that classification on edited design sets is worth further examination. It seems that unlimited editing and "all k-NN" editing in particular could result in design sets of a very desirable structure—essentially representing nonoverlapping probability distributions. These edited design sets would be very useful not only for k-NN classification but for any type of classification at all. Editing based on the k-NN rule could thus be a very desirable method of preprocessing before classification. It is hoped that the reported experiments will stimulate further research into the presented problem.

Appendix I

It will be shown that if a hypothesis on the asymptotic behavior of probability densities is true, then repeated editing of a very large design set generates eventually a design set which allows almost optimal classification by k-NN. We will say that the problem is well behaved if $P(C(i)/x',N)$ can be replaced by $P(C(i)/x,N)$ for M and all required values of N. Here M is the number of samples in design set D, $D(N)$ is the design set obtained from the original design set D by N editings, x' is the nearest neighbor of x in $D(N)$, and $P(C(i)/x,N)$ is the probability that $x \in D(N)$ belongs to class $C(i)$. It is shown in Wilson's paper that for a very general class of distributions the above replacement is legal for $D(1)$.

Theorem

Let the problem be well behaved. Let $\varepsilon > 0$ be given. Then there exists an integer m such that $i > m$ implies

$$R(i) \rightarrow \bar{R}(i) < R^* + \varepsilon$$

with probability 1, when the number of samples in D approaches infinity. $\bar{R}(i)$ is the asymptotic expected error.

Proof

Let $P(C(i)/x,N)$ be the probability that $x \in D(N)$ belongs to class $C(i)$. $x \in D(N+1)$ is in class $C(1)$ if and only if it is in class $C(1)$ in $D(N)$ and the majority of k nearest neighbors in $D(N)$ is from $C(1)$:

$$\begin{aligned} P(C(1)/x, N+1) = CN & \\ & \cdot P(C(1)/x,N) \sum_{j>k1} \binom{k}{j} P^j(C(1)/x',N) \\ & \cdot P^{k-j}(C(2)/x',N) \end{aligned}$$

(with $k1 = [k/2]$ the integer part of $k/2$, k odd and CN a normalizing constant).

If probabilities $P(C(i)/x,N)$ are well behaved, we can replace $P(C(i)/x',N)$ by $P(C(i)/x,N)$. This means that if

$$P(C(1)/x,N) = \alpha(x,N) \cdot P(C(2)/x,N) \tag{3}$$

then

$$\begin{aligned} \alpha(x, N+1) &= \frac{P(C(1)/x, N+1)}{P(C(2)/x, N+1)} \\ &= \alpha(x,N) \frac{\sum_{j>k1} \binom{k}{j} \alpha^j(x,N) \cdot P^k(C(2)/x,N)}{\sum_{j\le k1} \binom{k}{j} \alpha^j(x,N) \cdot P^k(C(2)/x,N)} \end{aligned}$$

so that

$$\alpha(x, N+1) = \alpha(x,N) \frac{\sum_{j>k1} \binom{k}{j} \cdot \alpha^j(x,N)}{\sum_{j\le k1} \binom{k}{j} \cdot \alpha^j(x,N)}. \tag{4}$$

This means that if $\alpha(x,2) = \alpha(x,1)$ then

$$\beta = \alpha(x,1) = \alpha(x,2) = \cdots \tag{5}$$

(according to (4)). For (5) to hold the following must be satisfied

$$\delta(\beta) = \frac{\sum_{j>k1} \binom{k}{j} \beta^j}{\sum_{j\le k1} \binom{k}{j} \beta^j} = 1$$

or

$$\gamma(\beta) = \sum_{j>k1} \binom{k}{j} \beta^j - \sum_{j\le k1} \binom{k}{j} \beta^j = 0. \tag{6}$$

However,

$$\gamma(\beta) = \sum_{j\le k1} \binom{k}{j} (\beta^{k-j} - \beta^j),$$

and this equation has root $\beta = 1$. For $\alpha > 1$, $\gamma(\alpha) > 0$ and for $\alpha < 1$, $\gamma(\alpha) < 0$ so that $\beta = 1$ is only real positive root of (6). (We are interested only in real positive roots.) Since we are interested only in the value of N and not parameter x at this point, let us write α_N instead of $\alpha(x,N)$ to make notation simpler.

The fact that (6) has exactly one positive root means that if $\alpha_1 > \beta$ then

$$\alpha_1 < \alpha_2 < \alpha_3 < \cdots$$

and if $\alpha_1 < \beta$ then

$$\alpha_1 > \alpha_2 > \alpha_3 > \cdots.$$

This can be seen from (4). The series $\{\alpha_i\}$ thus has a limit $A(\alpha_1)$. If $\alpha_1 > 1$ then $A(\alpha_1) = \infty$ and if $\alpha_1 < 1$ then $A(\alpha_1) = 0$: let $\alpha_1 > 1$. Let us assume that $A(\alpha_1) < \infty$. Since $A(\alpha_1)$ must be positive (> 1), $\delta(\alpha_i)$ also has a limit which is $\delta(A(\alpha_1))$. This means that

$$\begin{aligned} A(\alpha_1) &= \lim_{N\to\infty} \alpha_N = \lim_{N\to\infty} \alpha_{N+1} \\ &= \lim \alpha_N \cdot \delta(\alpha_N) = A(\alpha_1) \cdot \delta(A(\alpha_1)) \end{aligned}$$

so that

$$\delta(A(\alpha_1)) = 1.$$

However, this implies that $A(\alpha_1) = 1$ which is impossible. This means that $A(\alpha_1)$ must be ∞. Similarly $A(\alpha_1) = 0$ for $\alpha_1 < 1$.

Let us now partition S into $S(0)$ (part of S where $p_1q_1(x) = p_2q_2(x)$), X and Y. X contains all points in which the probability of incorrect assignment is bounded by a fixed number $B(1)$: $0 < B(1) < 0.5$ from above. Y contains all the remaining points bu̧t is small. Let, for a given $\varepsilon > 0$, Y be such that the expected error on Y does not exceed ε:

$$\mathop{\mathsf{E}}_{Y} r(x,1) \le \varepsilon$$

and probability of nonoptimal assignment is at least $B(1)$ everywhere in Y. (For given distributions, $B(1)$ is thus given by the value of ε.) Such partitioning is always possible if we choose $B(1)$ sufficiently close to 0.5.

Let us now denote the probability of assignment in agreement with the optimal rule by $S(x,N)$ (for point $x \in D(N)$) and

$$\bar{S}(x,N) = 1 - S(x,N).$$

(We thus have $S(x,N) > B(1)$ in Y and $S(x,N) \le B(1)$ in X.) For the calculation of the expected error of classification we thus have

$$r(x,N) = pq(x)\bar{S}(x,N) + \bar{p}\bar{q}(x)S(x,N)$$

where the barred probabilities are those associated with classes chosen by the optimal rule and the unbarred with the complementary class. We have

$$\begin{aligned} R(N) &= \mathop{\mathsf{E}}_{S} r(x,N) = \int_S r(x,N)\,dx \\ &= \int_S pq(x)\bar{S}(x,N)\,dx + \int_S \bar{p}\bar{q}(x)S(x,N)\,dx \\ &\le \int_S \bar{p}\bar{q}(x)\,dx + \int_{X\cup Y} pq(x)\bar{S}(x,N)\,dx \\ &= R^* + R'(N) \end{aligned}$$

where

$$R'(N) \triangleq \int_{X \cup Y} pq(x)\bar{S}(x,N)\, dx \leq B(N) \int_X pq(x)\, dx + \varepsilon$$
$$= B(N) \cdot C(X) + \varepsilon.$$

Here $C(X)$ is a finite constant whose value depends upon the definition of set X. $B(N)$ is related to series $\{\alpha_i\}$ in the following way. $B(1)$ defines both an upper bound on $S(x,N)$ on X and with the corresponding lower bound on $S(x,N)$ the first term of the series: α_1. Since $B(1) < 0.5$, $\alpha_1 < 1$. This means that $\alpha_N \to 0$ for $N \to \infty$. Since $S(x,N) + \bar{S}(x,N) = 1$ this implies that $B(N) \to 0$ with $N \to \infty$ and so

$$\lim_{N \to \infty} R'(N) = \lim B(N) \cdot C(X) + \varepsilon = \varepsilon$$

This completes the proof of the theorem.

Appendix II

The proof given in Appendix I will be used to show that the rate of convergence from the original distributions to the disjoint ones is very fast. Furthermore, the "all NN" rule will be shown to be converging much faster than the ordinary k-NN rule.

From the derivation in Appendix I we have

$$\alpha_{N+1} = \alpha_N \frac{\sum_{j>k1} \binom{k}{j} {\alpha_N}^j}{\sum_{j\leq k1} \binom{k}{j} {\alpha_N}^j} \leq \frac{\alpha_N \cdot \alpha_N^{k1+1} \sum_{j>k1} \binom{k}{j}}{{\alpha_N}^{k1} \sum_{j \leq k1} \binom{k}{j}} = {\alpha_N}^2$$

which holds for $\alpha_N < 1$. This shows that the speed of convergence is exponential, and thus very few iterations will considerably improve the suitability of D for classification.

Let us now consider "all NN" classification. In this case $x \in C(1) \cap D(N+1)$ if and only if the majority of its 3 and 5 and $\cdots k$ nearest neighbors from $D(N)$ are from $C(1)$. Using

$$k1(l) = [l/2]$$

and

$$p(k) = \left[\frac{k+1}{2}\right]$$

for k odd we have

$$\alpha_{N+1} = \alpha_N \frac{\prod_{l=1,3,\cdots,k} \sum_{j>k1(l)} \binom{l}{j} {\alpha_N}^j}{\prod_{l=1,3,\cdots,k} \sum_{j\leq k1(l)} \binom{l}{j} {\alpha_N}^j} \leq \alpha_N^{2p(k)}.$$

This shows that theoretically "all NN" elimination converges much faster than ordinary k-NN elimination, particularly for larger values of k.

The approximation of α_{N+1}/α_N used above is good for small values of k. For larger values of k the actual ratio is, in fact, smaller (for $\alpha_N < 1$) and convergence faster. This means that for larger values of k convergence is faster than for smaller values.

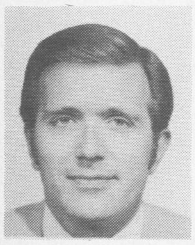
References

[1] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, Wiley, 1973.

[2] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 21–27, 1967.

[3] D. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Trans. Syst., Man, Cybernetics*, vol. SMC-2, pp. 408–421, 1972.

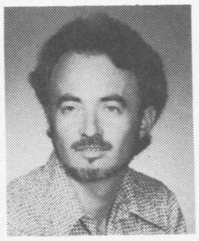
Contributors



John J. Allan, III, (M'69) was born in White Plains, N.Y., in 1937. He received the B.M.E. degree from The Ohio State University in 1960, and the M.S.E. and Ph.D. degrees from The University of Michigan in 1964 and 1968, respectively.

From 1958 to 1963 he was working with various industries such as the National Electric Coil Company, Kimble Glass Division of Owens–Illinois, and the Holophane Company in designing automated equipment and using and designing numerically-controlled machines. He joined the Department of Mechanical Engineering at The University of Texas at Austin in 1968 following the completion of his doctorate in the area of computer-based systems for engineering design. In addition to his regular duties as a faculty member, he is director of the Computer Applications Laboratory.

Dr. Allan is a technical and educational consultant to several corporations, a Registered Professional Engineer in Ohio and Texas, and is a member of ASME, ASEE, Sigma XI, The New York Academy of Sciences, and AAUP. He is Vice-Chairman of Working Group 5.2, Computer Aided Design, International Federation for Information Processing (IFIP), Chairman of the Research in CAD Committee, American Society of Mechanical Engineers (ASME), and an international traveling speaker for IEEE. He has published over 40 papers, a chapter in a 1973 book entitled *Computer-Aided Design*, an IEEE short course workbook, and is editor of the IFIP CAD glossary. He is listed in *American Men of Science*, *International Scholars Directory*, *Personalities of the South*, *Men of Achievement*, *1974*, and *Who's Who in the South and Southwest*. He has lectured in 18 foreign countries.



Yaakov Bar-Shalom (S'63–M'66) was born on May 11, 1941. He received the B.S. (cum laude) and M.S. degrees from the Technion—Israel Institute of Technology, in 1963 and 1967 and the Ph.D. degree from Princeton University in 1970, all in electrical engineering.

During 1963–1966 he served in the Israel Defense Army as an Electronics Engineer. From 1966 to 1967 he was a Teaching Assistant in the Department of Electrical Engineering of the Technion. Between 1967 and 1970 he was at Princeton University, first as a Research Assistant, then Research Associate. During the academic year 1968–1969 he was appointed Charles Grosvenor Osgood Fellow in recognition for