Appendix
Bayes Method for Known Number of Objects in Each Class

Consider a set of N objects, consisting of N_i objects originating from population A_i, $i=1,\cdots,k$, so that $N=\sum_{i=1}^{k}N_i$. In order to classify the set of N objects, we measure a variable with pdf f_i for objects originating from population A_i. Let $x_1,x_2,\cdots,x_N$ be the measured values of the N objects, which are assumed to be independent. Now consider the N-tuple $G_j=(g_{1j},g_{2j},\cdots,g_{ij},\cdots,g_{Nj})$, indicating that the object with measured value x_i originates from population $A_{g_{ij}}$, with $g_{ij}\in\{1,2,\cdots,k\}$ for $i=1,2,\cdots,N$. The number of elements in G_j with value l is equal to N_l $(l=1,2,\cdots,k)$. A possible classification of the set of N objects can be represented by the N-tuple $C=(c_1,c_2,\cdots,c_i,\cdots,c_N)$, where $c_i\in\{1,2,\cdots,k\}$, indicating that the object with measured value x_i is allocated to class c_i (there is no restriction on the number of elements in C with a specific value). We find for the mean number of misallocations, given the measured values $x_1,x_2,\cdots,x_N$, according to C

$$E(m)=\sum_{j=1}^{N_t}\Delta(C,G_j)\frac{f_{g_{1j}}(x_1)f_{g_{2j}}(x_2)\cdots f_{g_{Nj}}(x_N)P(G_j)}{\sum_{i=1}^{N_t}f_{g_{1i}}(x_1)f_{g_{2i}}(x_2)\cdots f_{g_{Ni}}(x_N)P(G_i)} \tag{5}$$

where $\Delta(C,G_j)$ is the number of misallocations given G_j, N_t equals the number of permutations of G, and $P(G_j)=1/N_t$. Now the classification C has to be chosen such that $E(m)$ is a minimum. After some reduction of (5) that classification criterion becomes: choose C so that

$$\sum_{j=1}^{N_t}\Delta(C,G_j)f_{g_{1j}}(x_1)f_{g_{2j}}(x_2)\cdots f_{g_{Nj}}(x_N) \tag{6}$$

is a minimum. Putting $\Delta(C,G_j)=\sum_{i=1}^{N_t}\delta(c_i,g_{ij})$, with $\delta(c_i,g_{ij})=0$ for $c_i=g_{ij}$ and $\delta(c_i,g_{ij})=1$ for $c_i\neq g_{ij}$, the c_i can be chosen independently, so that the classification criterion becomes: choose c_i $(i=1,2,\cdots,k)$ so that

$$\sum_{j=1}^{N_t}\delta(c_i,g_{ij})f_{g_{1j}}(x_1)f_{g_{2j}}(x_2)\cdots f_{g_{Nj}}(x_N) \tag{7}$$

is a minimum. With this latter classification criterion it is possible to allocate objects sequentially. However, the summation has to be done $N_t=N!/(N_1!N_2!\cdots!N_k!)$ times so that for large values of N and (especially) k, this classification criterion is not currently useful in practice because of its long computation time.

Acknowledgment

The author wishes to thank Prof.dr.ir. C. J. D. M. Verhagen, Dr.ir. F. C. A. Groen, and Dr.ir. R. P. W. Duin for their constructive criticism, and all the members of the group who have contributed to this correspondence.

References

[1] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley-Interscience, 1973.

[2] R. E. Slot, "On the profit of taking into account the known number of objects per class in classification methods," Dept. of Appl. Phys., Delft Univ. of Tech., Delft, Netherlands, *Internal. Rep.*, 1974.

[3] F. C. A. Groen, "Analysis of DNA based measurement methods applied to human chromosome classification," Thesis, Delft, 1977.

[4] M. L. Mendelsohn, "DNA-content and DNA-based centromeric index of the 24 human chromosomes," *Science*, vol. 179, pp. 1126–1129, March 1973.

The Condensed Nearest Neighbor Rule Using the Concept of Mutual Nearest Neighborhood

K. CHIDANANDA GOWDA AND G. KRISHNA

Abstract—**A two-stage iterative algorithm for selecting a subset of a training set of samples for use in a condensed nearest neighbor (CNN) decision rule is introduced. The proposed method uses the concept of mutual nearest neighborhood for selecting samples close to the decision line. The efficacy of the algorithm is brought out by means of an example.**

I. Introduction

The nearest neighbor (NN) rule [1]–[3] assigns an unclassified sample to the same class as the nearest of the N stored labeled samples of the training set. The rule is simple, yet powerful, and with an unlimited number of samples, the risk in making an NN decision is never worse than twice the Bayes risk [1]. But, as all the labeled samples of the training set must be searched to classify a test sample, the NN method imposes large storage and computational requirements.

The condensed nearest neighbor (CNN) rule [4], [9] retains the basic approach of the NN rule but uses only a subset of the training set of samples. This subset, when used as a stored reference set for the NN decision rule, correctly classifies all the samples belonging to the original training set. As the CNN method chooses samples randomly, internal rather than boundary samples are occasionally retained. Gates [5] has proposed the reduced NN modification of the CNN algorithm. Swonger [6] describes an iterative condensation algorithm for selecting a consistent subset of samples for use in a CNN decision rule. Ritter *et al.* [7] introduce an algorithm for a selective nearest neighbor decision rule. Tomek [8] presents two modifications of CNN by growing the condensed set using only samples close to the decision boundary. Chidananda Gowda and Krishna [10]–[12] have introduced the concept of mutual nearest neighborhood and a new similarity measure called the mutual neighborhood value (MNV). A two-stage algorithm of a modified CNN rule, making use of the MNV, will be described in the sequel.

II. Concepts and Definitions

We define the mutual neighborhood value between any two samples of a set as the sum of the conventional nearest neighbor ranks of these two samples with respect to each other.

Let $X_1,X_2,\cdots,X_N$ be a set of N L-dimensional vectors called samples, where the X_i's take values in a metric space upon which is defined a metric d. Let X_j be the mth nearest neighbor of X_i, and X_i be the nth nearest neighbor of X_j. Then the MNV between X_i and X_j is defined as $m+n$. That is, MNV $(X_i,X_j)=m+n$ where $m,n\in\{0,1,2,\cdots,N-1\}$ where 0 is used when $i=j$.

On the other hand, if we consider only k-nearest neighbors of each sample point, then if either X_i or X_j, or both, are not found in each other's k-nearest neighborhood, we say that X_i and X_j do not belong to the mutual neighborhood.

III. A Modified CNN Algorithm

A two-stage algorithm for selecting a subset of samples for use in a modified condensed nearest neighbor decision rule is described. Here ORDER, STORE, and SIFT are three storage bins.

Manuscript received February 21, 1978; revised June 20, 1978.

K. Chidananda Gowda is with the Department of Electrical Engineering, Sri Jayachamarajendra College of Engineering, Mysore-570 006, India.

G. Krishna is with the School of Automation, Indian Institute of Science, Bangalore 560012, India.

0018-9448/79/0700-0488$00.75 ©1979 IEEE

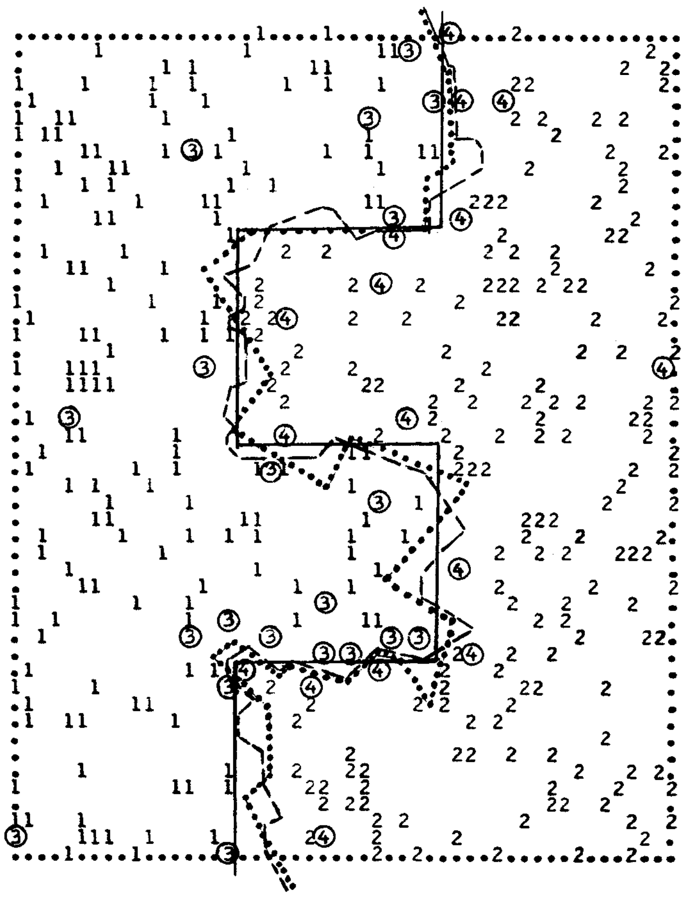

Fig. 1. Samples selected by Hart's method.

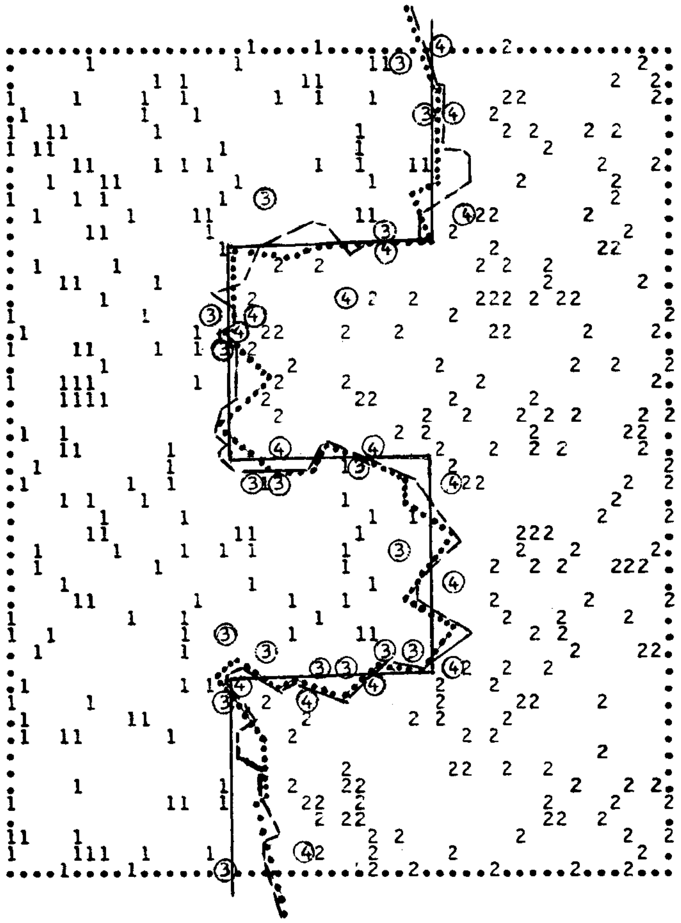

Fig. 2. Samples selected by Tomek's method.

Stage 1

1) For each sample X of the training set T, find the nearest neighbor Y from the opposite class. Subsequently, with respect to Y, considering only samples belonging to the opposite class, find the NN rank J of X. Now the MNV of X with respect to Y is MNV $(X,Y)=1+J$. This value of MNV will be associated with the sample X alone and not with Y. Also record the Euclidean distance d between X and Y, and associate it with X. Samples that are near the decision boundary will have low values of MNV and d.

2) Using the results from 1), order the N samples according to MNV in ascending order. If the MNV's of some of the samples are identical, order such samples according to distances d, in ascending order. Store this ordered set in ORDER.

3) The first sample of ORDER is placed in STORE.

4) The next sample in ORDER is classified by the NN rule, using the samples that are present in STORE. If the classification is wrong, add that sample to STORE.

5) Step 4) is repeated till all the samples in ORDER are tested.

6) After one pass through ORDER, apply steps 4) and 5) to the samples retained in ORDER. Repeat this procedure until there are no transfers of samples from ORDER to STORE in a pass. The present contents of STORE constitute the first modified condensed training set.

Stage 2

When the deletion of a sample in the condensed subset produces no change in the classification of any member of the complete training set, the deleted sample may be excluded from the condensed set [5]. This idea is used to make a further reduction in the number of samples constituting the modified condensed set.

7) A sample Z of STORE (on completion of step 6)) is placed in SIFT.

8) All the samples in ORDER are classified by the NN rule using the samples that are now present in STORE. If there is any misclassification, transfer Z back to STORE, else retain it in SIFT.

9) Steps 7) and 8) are repeated for all the samples in STORE. The final contents of STORE constitute the second modified condensed training set.

IV. Experimental Results

An example described by Ritter [7] has been repeated here to bring out the efficacy of the proposed method. A two-class problem consisting of two-dimensional uniform distributions is generated according to the decision boundary (solid lines) shown in Figs. 1–4. 400 samples, approximately 200 in each class, are generated from the distributions. In Figs. 1–4, samples marked 1 or 3 belong to class 1, and samples marked 2 or 4 belong to class 2 in the training set. Samples marked 3 belong to class 1, and samples marked 4 belong to class 2 in the condensed training set.

When the CNN algorithm proposed by Hart [4] is applied to the data, it terminates giving a condensed subset containing 36 samples. The samples belonging to this subset are marked 3 (class 1) and 4 (class 2) in Fig. 1. It can be observed that some of the samples belonging to the condensed training set are not at all near the decision boundary.

When the second modified CNN algorithm proposed by Tomek [8] is used on the data, the algorithm ends giving a condensed set with 34 samples as shown in Fig. 2.

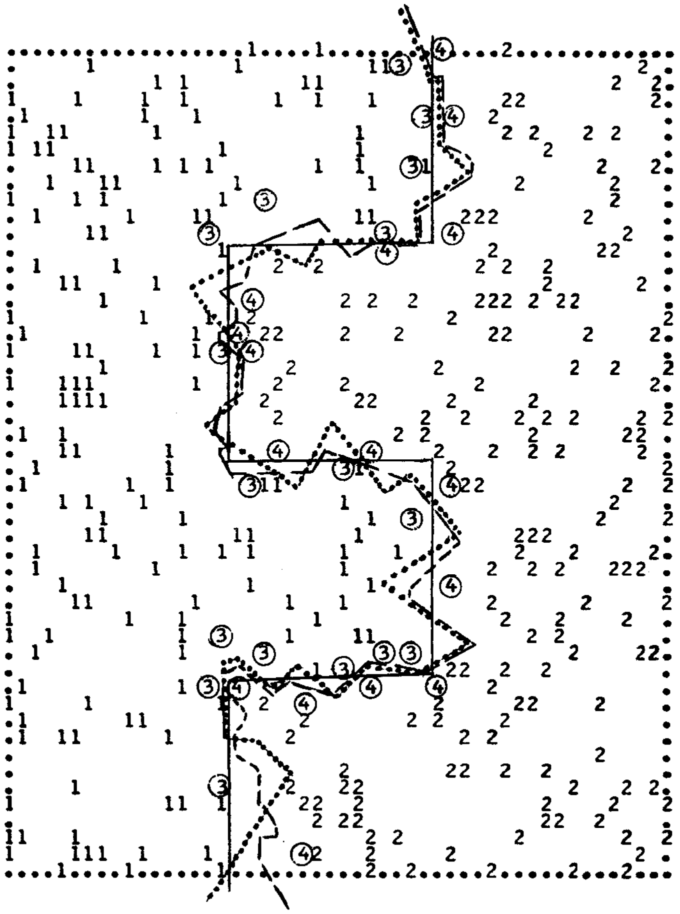

Fig. 3. Samples selected by proposed method (stage 1).

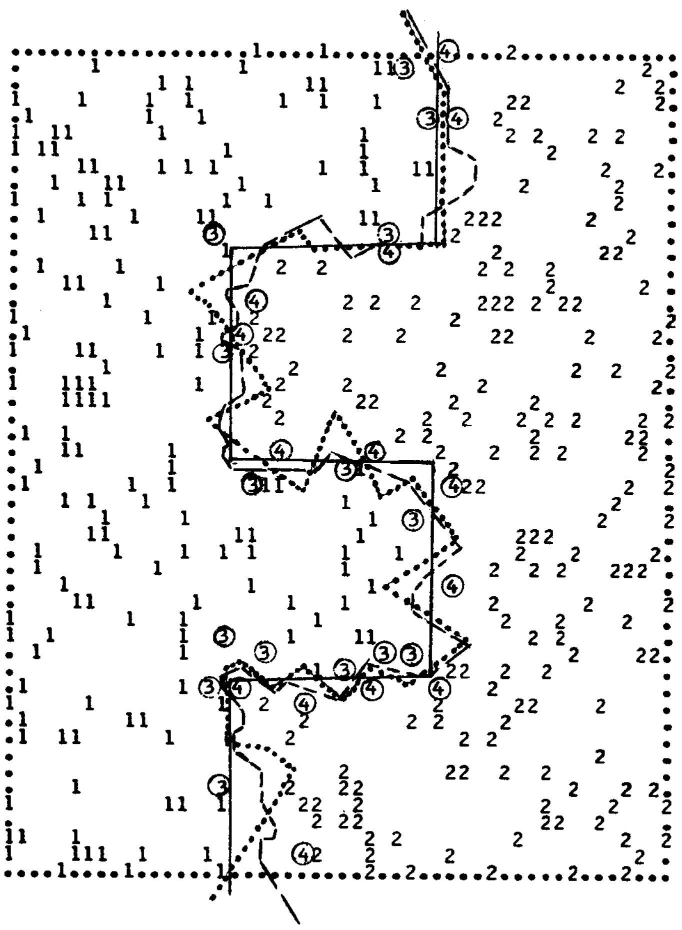

Fig. 4. Samples selected by proposed method (stage 2).

The first stage of the algorithm proposed in this paper yields a condensed set with 33 samples. This is illustrated in Fig. 3. The application of the second stage of our algorithm further reduces the training set to a subset containing 29 samples. This result is illustrated in Fig. 4.

For the example considered, the computer time for Hart's algorithm is 83 s, for Tomek's algorithm 560 s, and for the proposed algorithm 951 s. Hart's algorithm is computationally very economical, but its main drawback is that it retains samples that are not near the decision boundary. The proposed algorithm generates a condensed set, comprising samples near the decision boundary, which is the smallest for the example considered.

In Figs. 1–4, the boundary drawn in solid lines is used to generate 400 samples of the training set belonging to the two classes. Using this training set of samples and the NN rule, the piecewise linear boundary, as shown in dashed lines, is drawn. The piecewise linear boundaries resulting from the condensed training set of each method and the NN rule are drawn in dotted lines. These boundaries indicate, to some extent, the distortions introduced by the different training sets.

Conclusion

The concept of mutual nearest neighborhood is used to obtain a modified condensed training set. The proposed method prevents the retention of interior samples in the condensed set by seeking to add samples close to the decision boundary. The efficacy of the procedure is illustrated by an example.

Acknowledgment

The authors wish to thank Prof. K. S. Prabhu, Chairman of the Department of Electrical Engineering, Indian Institute of Science, Bangalore, for providing necessary facilities and giving encouragement from time to time. K. Chidananda Gowda is grateful to the Government of India and the Government of Karnataka for awarding him a Q.I.P. fellowship.

References

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

[2] T. M. Cover, "Estimation by the nearest neighbor rule," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 50–55, May 1968.

[3] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[4] P. E. Hart, "The condensed nearest neighbor rule," *IEEE Trans. Inform. Theory* (Corresp.), vol. IT-14, pp. 515–516, May 1968.

[5] G. W. Gates, "The reduced nearest neighbor rule," *IEEE Trans. Inform. Theory* (Corresp.), vol. IT-18, pp. 431–433, May 1972.

[6] C. W. Swonger, "Sample set condensation for a condensed nearest neighbor decision rule for pattern recognition," presented at the Int. Conf. Frontiers of Pattern Recognition, Honolulu, Hawaii, Jan. 18–20, 1971.

[7] G. L. Ritter, H. B. Woodruff, S. R. Lowry, and T. L. Isenhour, "An algorithm for a selective nearest neighbor decision rule," *IEEE Trans. Inform. Theory* (Corresp.), vol. IT-21, pp. 665–669, Nov. 1975.

[8] I. Tomek, "Two modifications of CNN," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-6, pp. 769–772, Nov. 1976.

[9] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York: Academic, 1972.

[10] K. Chidananda Gowda and G. Krishna, "Nonparametric clustering using the concept of mutual nearest neighborhood," Rep. No. EE/43, Dep. Elec. Eng., Indian Institute of Science, Bangalore, India, Apr. 1977.

[11] ——, "Agglomerative clustering using the concept of mutual nearest neighborhood," *Pattern Recognition*, vol. 10, no. 2, pp. 105–112, 1978.

[12] ——, "Clustering and unsupervized learning using the concept of mutual nearest neighborhood," Rep. No. EE/49, Dep. Elec. Eng., Indian Institute of Science, Bangalore, India, Jan. 1978.