The Distance-Weighted k-Nearest-Neighbor Rule

SAHIBSINGH A. DUDANI

***Abstract*—Among the simplest and most intuitively appealing classes of nonprobabilistic classification procedures are those that weight the evidence of nearby sample observations most heavily. More specifically, one might wish to weight the evidence of a neighbor close to an unclassified observation more heavily than the evidence of another neighbor which is at a greater distance from the unclassified observation. One such classification rule is described which makes use of a neighbor weighting function for the purpose of assigning a class to an unclassified sample. The admissibility of such a rule is also considered.**

I. Introduction

For a given classification problem, a designer might find himself in one of two extreme situations with regard to *a priori* knowledge about the problem. In one instance, he may have complete statistical knowledge of the underlying joint distribution for observation ρ and category η. In such a case, a Bayes classification procedure would yield an optimum design and the corresponding minimum (Bayes) probability of classification error. In the second case, the available information on the problem may consist only of a collection of n correctly classified samples $(\rho_1,\eta_1),(\rho_2,\eta_2),\cdots,(\rho_n,\eta_n)$. In this case, no classification procedure exists which would be optimum with respect to all of the possible underlying joint distributions that might have yielded the given sample set.

Among the simplest and most intuitively appealing classes of nonprobabilistic classification procedures are those that weight the evidence of nearby sample observations most heavily. The first formulation of a rule of the nearest-neighbor type appears to have been made by Fix and Hodges [1]. They investigated a rule called the k-nearest-neighbor rule, which assigns to an unclassified sample that class most heavily represented among its k nearest neighbors. Cover and Hart [2] studied most extensively the properties of this rule, including the lower and upper bound on the probability of error. It has been demonstrated that when k and n tend to infinity in such a manner that $k/n \rightarrow 0$, the risk of such a rule approaches the Bayes risk.

In this correspondence, a nonprobabilistic classification procedure of nearest-neighbor type is described which makes use of a neighbor weighting function for the purpose of assigning a class to an unclassified sample.

II. Decision Rule

It is reasonable to assume that observations which are close together (according to some appropriate metric) will have the same classification. Furthermore, it is also reasonable to say that one might wish to weight the evidence of a neighbor close to an unclassified observation more heavily than the evidence of another neighbor which is at a greater distance from the unclassified observation. Therefore, one would like to have a weighting function which varies with the distance between the sample and the considered neighbor in such a manner that the value decreases with increasing sample-to-neighbor distance. One such weighting function is employed below in defining a "distance-weighted k-nearest-neighbor rule."

Let each pattern ρ_i in the training set (collection of correctly classified samples) be associated with a category number η_i, where $\eta_i \in \{1,2,\cdots,R\}$. When an unknown pattern ρ' is to be classified, the k nearest neighbors of ρ' (according to a suitable metric) are found among the given samples constituting the training set. Let these k nearest neighbors of ρ', with their associated category numbers, be given by (ρ_j',η_j), $j = 1,\cdots,k$. The neighbors (ρ_j',η_j), $j = 1,\cdots,k$ are ordered so that ρ_j' is the nearest and ρ_k' the farthest from the unknown sample ρ'. Also, let the corresponding distances of these neighbors from the unknown pattern ρ' be given by d_j, $j = 1,\cdots,k$. A weight w_j attributed to the jth nearest neighbor can be defined as

$$w_j = \begin{cases} \dfrac{d_k - d_j}{d_k - d_1}, & d_k \neq d_1 \\ 1, & d_k = d_1 \end{cases} \tag{1}$$

It should be noted that the value of w_j varies from a maximum of 1 for a nearest neighbor down to a minimum of zero for the most distant of the k neighbors.

Having computed the weights w_j, the distance-weighted k-nearest-neighbor rule then assigns the unknown pattern ρ' to that class for which the weights of the representatives among the k nearest neighbors sum to the greatest value. It can be seen from the definition of the weighting function given in (1) that it is worthy of consideration only for values of k greater than 3.

If the number of training samples n is very large compared with the number of nearest neighbors k considered, then it is reasonable to expect that the results obtained through the two rules—the distance-weighted k-nearest-neighbor rule and the simple majority k-nearest-neighbor rule—will be comparable to each other. However, it is the author's belief that use of the distance-weighted k-nearest-neighbor rule with training sample sets of small or moderate size will yield smaller probabilities of error, at least for certain classes of distributions.

III. Admissibility of Decision Rule for Small and Moderate Size Sample Sets

The purpose of this section is to show that, considering the class of nearest-neighbor rules, the distance-weighted k-nearest-neighbor rule is admissible for small and moderate size training sample sets. That is, it will be shown here that for at least one arbitrarily chosen example of a small training sample set, drawn from a certain set of probability distribution functions, a lower probability of error P_e is obtained for the distance-weighted k-nearest neighbor rule than for a simple majority k-nearest-neighbor rule.

Consider a three-class problem with equal *a priori* class probabilities. Let the three classes have the bivariate distributions shown below.

Class 1: $p_1(\rho) = \frac{3}{5}N(\mu_{11},\Sigma_1) + \frac{2}{5}N(\mu_{12},\Sigma_1)$,

where

$$\mu_{11} = (3.0,3.0)$$

$$\mu_{12} = (7.0,7.0)$$

$$\Sigma_1 = \begin{vmatrix} 2.25 & 0.0 \\ 0.0 & 2.25 \end{vmatrix}.$$

Class 2: $p_2(\rho) = N(\mu_2,\Sigma_2)$,

where

$$\mu_2 = (4.0,6.0)$$

$$\Sigma_2 = \begin{vmatrix} 4.0 & 0.0 \\ 0.0 & 4.0 \end{vmatrix}.$$

Manuscript received May 21, 1975; revised October 17, 1975. The concept of a distance-weighted decision rule was originally introduced in the author's Ph.D. dissertation completed at the Ohio State University, Columbus, OH, with the support of the United States Air Force Office of Scientific Research under Grant No. AF-AFOSR-71-2048.

The author is with Hughes Research Laboratories, Malibu, CA 90265.

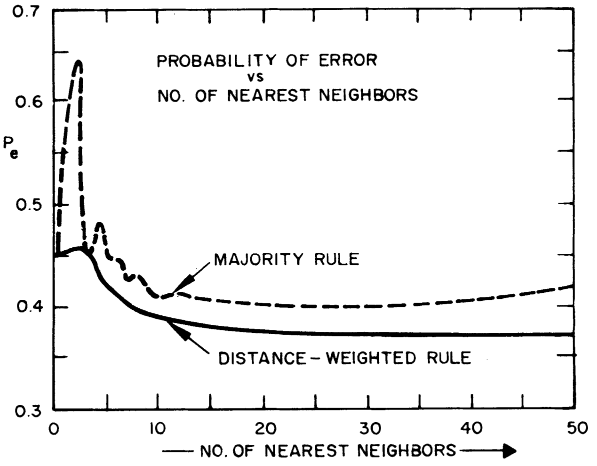

Fig. 1. Plots of probability of error with respect to k for the two nearest-neighbor type rules.

Class 3: $p_3(\rho) = N(\mu_3,\Sigma_3)$,

where

$$\mu_3 = (7.5,3.5)$$

$$\Sigma_3 = \begin{vmatrix} 9.0 & 0.0 \\ 0.0 & 9.0 \end{vmatrix}.$$

A Monte Carlo analysis was performed to estimate the probability of error for the above example for the case of a training set of 150 samples, including an equal number of samples from each of the classes. In order to avoid getting certain statistical deviations in the results of the analysis, a very large set of test samples of size 3000 was used. Also the training set of 150 was drawn independently six different times, and the classification test was performed using each of these training sets. A Euclidian metric was used here for obtaining the distances of neighbors. The plots of the estimates for probabilities of error with respect to k, the number of nearest neighbors employed in the two rules, appear in Fig. 1. It can be seen from these plots that the probability of error for the distance-weighted rule is strictly lower than that for the majority rule for values of k greater than three. Thus it can be said that the distance-weighted k-nearest-neighbor rule is admissible.

IV. Discussion

It should be noted here that the high value of probability of error for the majority rule in the case of $k = 2$ is a result of the occurrence of many ties. It is fair to say that at least one of the reasons for a lower probability of error for the distance-weighted k-nearest-neighbor rule than for the simple majority rule in the case of small training set sizes is due to the fact that the probability of ties taking place in the distance-weighted rule is strictly lower than that encountered in the case of the simple majority rule.

One other desirable feature of the distance-weighted rule may be pointed out here. For a simple majority rule, it is difficult to select a nearly optimum value of k to approach the lowest possible probability of error because of the nature of the variation of probability of error with the number of nearest neighbors k. The nature of this variation may be attributed to the fact that the possibility of ties taking place depends heavily on the value of k and on the number of classes present. Also, as k is increased beyond a certain value, which depends on the number of samples in the training set, the probability of error may begin to increase in certain cases. As an example, the use of majority rule for the classification of vocal utterance [3] showed that the recognition accuracy is higher when the classifier bases its decision on the class of the nearest neighbor rather than the most populous class among the nine nearest neighbors. These difficulties in selecting the value of k do not arise for the distance-weighted k-nearest-neighbor rule. Thus, for the distance-weighted rule, one could safely select a fairly large value of k without fear of the probability of error obtained being significantly greater than that which could be achieved with the optimum value of k. The plots of Fig. 1 support the above discussion.

Also, it should be mentioned here that application of the distance-weighted k-nearest-neighbor rule discussed in this correspondence to an automatic aircraft identification problem [4] yielded better results (lower probability of error) than application of the simple majority rule.

V. Examples of Other Weighting Functions

It was pointed out earlier that a weighting function employed by the distance-weighted k-nearest-neighbor rule should be such that it varies with the distance between the sample and the considered neighbor in such a manner that the value decreases with increasing sample-to-neighbor distance. An example of such a weighting function considered earlier for the distance-weighted k-nearest-neighbor rule has been given in Section III. Another example of a weighting function possessing the property mentioned above is shown below:

$$w_j = \frac{1}{d_j}, \qquad d_j \neq 0. \tag{2}$$

For this example of a weighting function, the relationship between the distance d_j and the corresponding weight w_j is such that the weights w_j take very large values for distances d_j close to zero, and thus reduce the corresponding classification algorithm in many cases to a simple nearest-neighbor rule.

An example of a weighting function that depends on the rank of the neighbor rather than its distance from the unknown sample is

$$w_j = k - j + 1. \tag{3}$$

Note that w_j takes only integer values from k to 1 in the above equation. This type of weighting function has a clear drawback of not being able to adjust the distribution of weights, depending on closeness of the neighbors to the unknown sample. However, the computation time involved for this weighting function is less than for the ones defined in (1) and (2).

VI. Conclusion

We have described here a k-nearest-neighbor decision rule which uses a distance function to weight the evidence of a neighbor close to an unclassified observation more heavily than the evidence of another neighbor which is at a greater distance from the unclassified observation. It was shown that for at least one arbitrary chosen example of a small training set, a lower probability of misclassification was obtained for the distance-weighted k-nearest-neighbor rule than for a simple majority k-nearest-neighbor rule. Based on the results shown in this correspondence and the use of this decision rule in an aircraft identification problem, it is the author's belief that use of the distance-weighted k-nearest-neighbor rule with training sample sets of small or moderate size will yield smaller probabilities of error (misclassification), at least for certain classes of distribution.

ACKNOWLEDGMENT

The concept of distance-weighted decision rules was first developed in conversations with Dr. Timothy Hawkins. His helpful suggestions and comments are gratefully acknowledged.

REFERENCES

[1] E. Fix and J. L. Hodges, Jr., "Discriminatory analysis, nonparametric discrimination: Consistency Properties," USAF School of Aviation Medicine, Randolph Field, TX, Project No. 21-49-004, Rep. No. 4, Contract No. AF41(128)-31, Feb. 1951.

[2] T. M. Cover and P. Hart, "The nearest neighbor decision rule," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 21–27, Jan. 1967.

[3] G. M. White and P. J. Fong, "k-nearest-neighbor decision rule performance in a speech recognition system," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-5, p. 389, May 1975.

[4] S. A. Dudani, "An experimental study of moment methods for automatic identification of three-dimensional objects from television images," Ph.D. dissertation, The Ohio State University, Columbus, Aug. 1973.

An Efficient Algorithm for Constructing Hierarchical Graphs

LAWRENCE YELOWITZ

Abstract—**An algorithm to delete redundant edges from a precedence graph is presented and proved correct. The algorithm is much more efficient than previous algorithms to perform the same task.**

I. INTRODUCTION

In this note we consider the problem of deleting all the redundant edges from a cycle-free digraph (precedence graph). This problem has been considered previously [1], [2], and has many applications, e.g., determining maximal parallelism in a system of concurrent tasks [3, sect. 2.2]. We adapt the notations and definitions of [1] here with some minor changes. The differences between the present note and [1] are the following.

1) The present algorithm takes its input in the form of a (square) adjacency matrix M, rather than in the form of a connectivity table. M can be processed immediately in its initial form. In particular it is not necessary to perform relabelling to transform M to upper-triangular form; such a relabelling is required in [1].

2) The present algorithm computes the nonredundant matrix M corresponding to M *in situ*, and returns its output in M. Thus, no additional storage is needed for building paths, etc.

In Section IV we compare the efficiency of the present algorithm with the efficiency of [1].

II. THE ALGORITHM

Let M be an $N \times N$ adjacency matrix. During the execution of the algorithm, a nonzero, nonunit element will be used, denoted simply by "*". To interpret the resulting matrix M at termination as the desired nonredundant matrix, simply interpret the "*" as zero. Put differently, $M(I,J) = 1$ at termination iff edge (I,J) is a nonredundant edge. By interpreting the "*" as unity, the resulting matrix at termination equals the transitive closure of the original matrix. Table I presents the algorithm in a "PL/1-like" programming language, and Fig. 1 presents the algorithm in flowchart form. The proof of correctness in Section III refers to step 6, which is found in Table I.

TABLE I
ALGORITHM FOR TRANSITIVE REDUCTION EXPRESSED IN "PL/1-LIKE" PROGRAMMING LANGUAGE

```
1.  DO K=1 TO N
2.    DO I=1 TO K-1, K+1 TO N
3.      IF M(I,K)≠0 THEN
4.        DO J=1 TO K-1, K+1 TO N
5.          IF M(K,J)≠0 THEN
6.            M(I,J)="*"
7.        END
8.    END
9.  END
```

III. PROOF OF CORRECTNESS OF THE ALGORITHM

To prove that the algorithm works correctly, the following two facts must be shown.

1) Every nonredundant unit element present initially in M remains as a unit element at termination.

2) Every redundant unit element present initially in M has been changed to a "*".

Several theorems are now presented to establish these facts.

Theorem 1: If G is a precedence graph, then the nonredundant graph corresponding to G is unique.

Proof: Using a technique known as topological sorting [5, sect. 2.2.3] it is possible to index the nodes of the graph such that $I < J$ for each edge (I,J) in the graph. Define the level of edge (I,J) to be $J - I$. Define an edge as essential if the edge must occur in the nonredundant graph. Each edge in level 1 is essential. An edge in level k is essential iff it is not implied by transitivity of edges in level p, $1 \le p < k$. Thus the set of essential edges is uniquely determined, and is in fact the precise set of edges comprising the nonredundant graph.

Theorem 2: At any point in the execution of the algorithm, if $M(I,J) =$ "*" then there is a path from I to J based on the current unit elements in M.

Proof: Let $T = \langle t1,t2,\cdots,tq\rangle$ be the ordered moments of time during execution of the algorithm when control resides in step 6, i.e., when $M(I,J)$ is assigned the value "*". For $t \in T$, let t' and t'' denote moments of time immediately preceding and and following t. We show that if Theorem 2 is true at t', then it remains true at t'', for all $t \in T$.

Theorem 2 is vacuously true at $t1'$.

Let $t \in T$, and consider the following three cases.

1) $M(I,J) =$ "*" at time t'. Thus, there is no change to the matrix, and Theorem 2 remains true at time t''.

2) $M(I,J) = 1$ at time t'.

a) We first show that Theorem 2 is true of the specific element $M(I,J)$ at t''. At t' there exists some index K, where $K \neq I$ and $K \neq J$, such that $M(I,K) = 1$ or "*" and similarly $M(K,J) = 1$ or "*". If both $M(I,K)$ and $M(K,J) = 1$ then the desired result for $M(I,J)$ follows immediately.

Suppose that $M(I,K)$ or $M(K,J) =$ "*". The only way Theorem 2 could fail for $M(I,J)$ at t'' is for every indirect path from I to K (or every indirect path from K to J), to contain edge (I,J). However, the cycle-free property of the graph prevents this from occurring.

Manuscript received October 13, 1974; revised November 10, 1975.

The author is with the Computer Science Department, University of Pittsburgh, Pittsburgh, PA 15260.