A Counter-Example to Tomek's Consistency Theorem for a Condensed Nearest Neighbor Decision Rule*

Godfried T. Toussaint

School of Computer Science
McGill University
Montreal, Canada H3A 2A7

Abstract

The *condensed nearest neighbor rule* (CNN) was proposed by Hart [Ha68] as a method to reduce the storage requirements of the original data set D for the efficient implementation of the nearest neighbor decision rule in pattern classification problems. Tomek [To76a] suggested two modifications of CNN in order to improve its performance. As a first step in Tomek's second method he computes a subset C of D, for subsequent use in CNN, and claims that C is *training-set-consistent*, i.e., that all data points in D are correctly classified by the nearest neighbor rule using C. In this note we provide a counter-example to this claim. We also analyze Tomek's algorithm in the context of more recent graph-theoretical condensing schemes.

1. Introduction

In the non-parametric classification problem we have available a set of n feature vectors taken from a collected data set of n objects (patterns) denoted by $\{\mathbf{X},\Theta\} = \{(X_1,\theta_1), (X_2,\theta_2),..., (X_n,\theta_n)\}$, where X_i and θ_i denote, respectively, the feature vector on the ith object and the class label of the ith object. The labels θ_i are assumed to be correct and are taken from the integers $\{1,2,...,M\}$, i.e., the patterns may belong to one of M classes. One of the most attractive non-parametric decision rules is the so-called *nearest-neighbor* rule (*NN*-rule) [CH67], [De81]. Let X be a new object (feature vector) to be classified and let $X_j^* \in \{X_1, X_2,..., X_n\}$ be the feature vector closest to X, where closeness is measured by, say, the Euclidean distance between X and X_j^* in $\mathbf{R}^d$. The nearest neighbor decision rule classifies the unknown object X as belonging to class θ_j^*.

Let $P_e^n\,(NN) = Pr\{\theta \neq \theta_j^*\}$ denote the resulting probability of misclassification (error), where θ is the true class of X, and let $P_e\,(NN)$ denote the limit of $P_e^n\,(NN)$ as n approaches infinity. It has been shown by Cover and Hart [CH67] that as n goes to infinity the asymptotic nearest neighbor error is bounded in terms of the optimal Bayes error P_e by:

Therefore the asymptotic probability of error of the nearest neighbor rule is close to optimal. Furthermore, with a suitable modification of the *NN*-rule (the so-called *k-NN*-rule) we can ob-

* This research was supported by grants NSERC-OGP0009293 and FCAR-93ER0291.

$$P_e \le P_e(NN) \le P_e\left[2 - M\left(\frac{P_e}{M-1}\right)\right]$$

tain a probability of error as close to optimal as desired if k is chosen appropriately. In proving the above result Cover and Hart [CH67] originally invoked some restrictions on the underlying distributions but more recently Devroye [De81] and Stone [St77] proved the above results for *all* distributions. These strong performance bounds, together with the transparent simplicity of the rule, make it very attractive. However, the apparent (but misguided) necessity to store all the data $\{\mathbf{X},\Theta\}$ and the resulting excessive computational requirements, have unjustly discouraged many researchers from using the rule in practice.

In order to combat the storage problem, and resulting computation, many researchers, starting with Hart [Ha68], proposed schemes for "condensing" the original data $\{\mathbf{X},\Theta\}$ (also referred to in the literature as "reducing," "thinning," "editing," "pre-processing" and "prototype selection") so that fewer feature vectors need be stored. We should point out that more recently a more standardized terminology is being applied to this type of operation. For example, Devijver and Kittler [DK82] make a distinction between, on the one hand, eliminating outliers and overlapping prototypes in an attempt to improve the classification error and, on the other hand, recovering the nearest neighbor decision boundaries with an ideally minimal subset of prototypes. The former they call *editing* and the latter *condensing*.

Denote the condensed subset of $\{\mathbf{X},\Theta\}$ by $\{\mathbf{X},\Theta\}_E$. At least a dozen other examples of condensing schemes have been proposed in the literature[Ri75], [To76a], [To76b], [Sw72], [GK79], [FP70], [Ul74], [Ga72], [Ch74], [FM84]. All these techniques have several weaknesses in common. For one, they are *sequential* in nature and the resulting $\{\mathbf{X},\Theta\}_E$ is a function of the *order* in which $\{\mathbf{X},\Theta\}$ is processed. Secondly they all attempt to obtain a condensed set that will determine only *approximately* the original decision boundary in $\mathbf{R}^d$ that is determined by $\{\mathbf{X},\Theta\}$. To this end they use heuristics which often complicate the algorithms, in some cases requiring a great deal of computation if a minimal-size condensed set is required, and generally result in rather involved procedures that are very difficult to analyze theoretically. Furthermore, it has been shown that obtaining minimal size condensed sets with some of these algorithms is in fact NP-complete [Wi92]. While some of the schemes [Ha68] result in a condensed set that is *training-set consistent* (i.e., $\{\mathbf{X},\Theta\}_E$ classifies all objects in $\{\mathbf{X},\Theta\}$ correctly), none of the above yield a condensed set which is *decision-boundary consistent* (i.e., $\{\mathbf{X},\Theta\}_E$ defines precisely the same decision boundary in $\mathbf{R}^d$ as $\{\mathbf{X},\Theta\}$). Thus with the above condensing schemes we have not only the disconcerting fact that $\{\mathbf{X},\Theta\}_E$ does not implement the originally intended decision boundary, but we do not even know the relationship that exists, if any, between the resulting $\{\mathbf{X},\Theta\}_E$ and one that is decision-boundary consistent. Finally in 1984 an condensing scheme based on the Voronoi diagram of $\{\mathbf{X},\Theta\}$ was discovered that was both *training-set consistent* and *decision-boundary consistent* [TBP84]. We should point out here that although the Voronoi diagrams used in [TBP84] were based on the Euclidean metric, this is not an inherent limitation of Voronoi-based condensing schemes. Indeed, not only can Voronoi diagrams with arbitrary Minkowski metrics be constructed [Au91] but with Laguerre metrics [IIM85] and even more abstract measures [Au91], [KL87].

Before describing Tomek's procedure we outline Hart's CNN rule since the former method is a simple modification of the latter. To simplify notation and for ease of comparison, from here on we adopt Tomek's notation. Therefore let D denote the original data set $\{\mathbf{X},\Theta\}$ and E the resulting condensed set $\{\mathbf{X},\Theta\}_E$. The basic idea is to pick an arbitrary point from D and place it in an originally empty set E. Then the remaining points in D are classified by the NN-rule using E and those that are classified incorrectly are added to E. This procedure is iterated until no more data points are transferred from D to E. The justification for the method is that if a point is miss-classified it probably lies close to the decision boundary and therefore should be kept. The following is Tomek's description of Hart's CNN rule. It should be noted that this pseudo-code description did not appear in Hart's paper where an English version is described. Furthermore, there appears to be a typographical error in Tomek's description. To correspond with Hart's CNN the last statement should read ***go to*** (c) and not ***go to*** (b).

PROCEDURE CNN

begin

(a) pass $\leftarrow$ 1,

(b) choose $x \in D$ randomly, $D(1) = D - \{x\}$, $E = \{x\}$,

(c) $D(\text{pass} + 1) = \varnothing$, count $\leftarrow$ 0,

(d) choose $x \in D(\text{pass})$ randomly, classify x by the NN-rule using E,

(e) ***if*** classification in (d) is correct,

  then $D(\text{pass} + 1) = D(\text{pass} + 1) \cup x$,

  else $E = E \cup x$, count $\leftarrow$ count + 1,

(f) $D(\text{pass}) = D(\text{pass}) - \{x\}$,

(g) ***if*** $D(\text{pass}) \neq \varnothing$ ***go to*** (d),

(h) ***if*** count = 0

  then Exit

  else pass $\leftarrow$ pass + 1, ***go to*** (b)

end

2. Tomek's Modified Condensed Nearest Neighbor Rule

Tomek argues that CNN keeps too many points that are not near the decision boundary because of its arbitrary initialization step. In order to combat this he proposes as his second modification of CNN a preliminary pass of D to select a special subset of D called C. Then his method

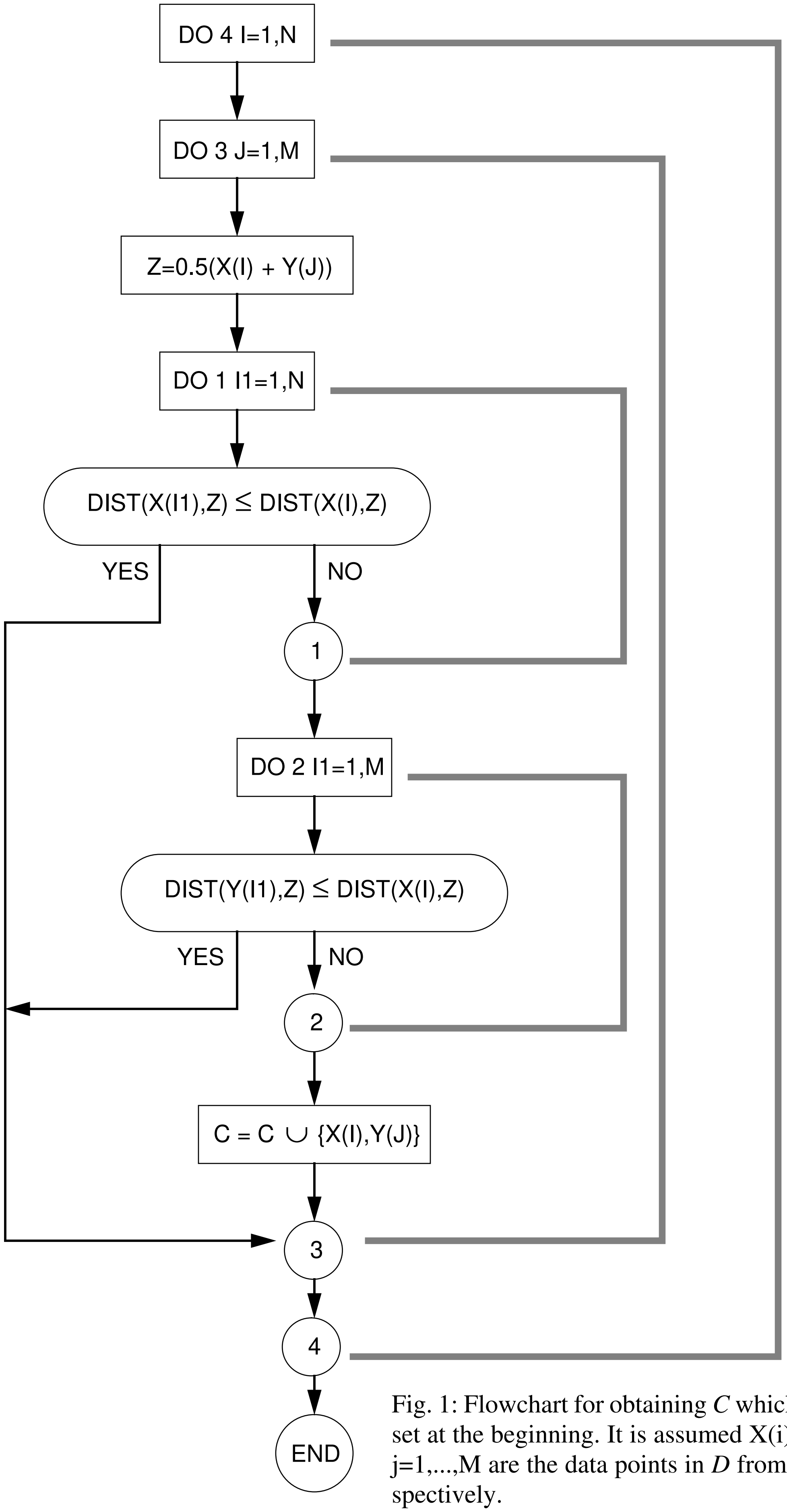

Fig. 1: Flowchart for obtaining *C* which is equal to the null set at the beginning. It is assumed X(i), i=1,...,N and Y(j), j=1,...,M are the data points in *D* from classes 1 and 2, respectively.

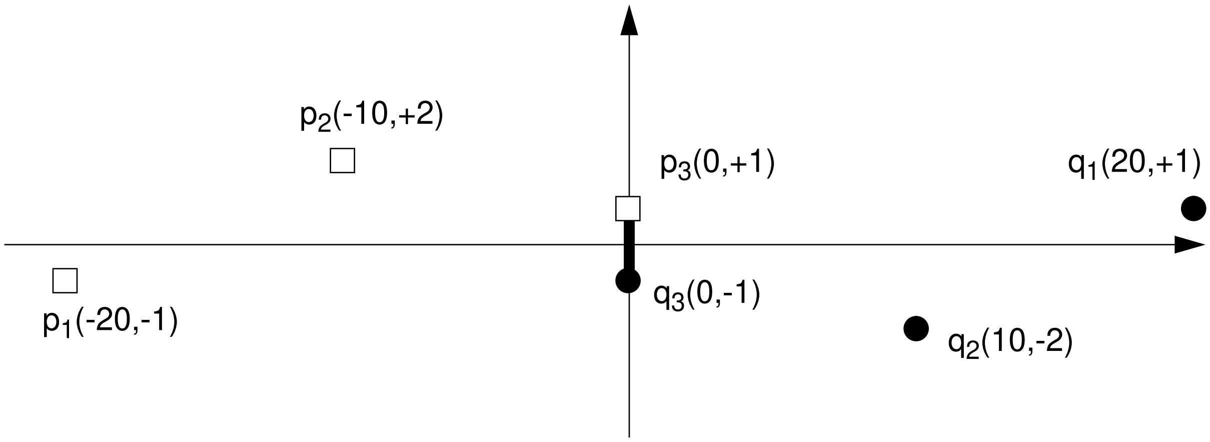

Fig. 2: A counter-example to Tomek's consistency theorem.

works in the same manner as CNN but instead of moving to *E* data points from the complete *D,* only data points from *C* are used. Tomek then describes an algorithm for computing *C* in the form of a FORTRAN IV flow chart which is duplicated in Fig. 1.

Tomek then states the following *training-set-consistency* theorem for which he includes an induction "proof" in the Appendix.

Theorem: All points in *D* are correctly classified by the NN rule using subset *C*.

3. The Counter-Example

We now show that the above claimed theorem is not valid. Consider a set *D* consisting of six data points distributed as in Fig. 2. Let $\{p_1, p_2, p_3\}$ denote points from class 1 and $\{q_1, q_2, q_3\}$ denote points from class 2. The points are shown with their corresponding *x* and *y* coordinates. Now consider the algorithm of Fig. 1 executing on this set *D*. The only pair of points (p_i, q_j) which satisfies both distance criteria tested in the innermost loops in Fig. 1, is (p_3, q_3) and hence the set $C = \{p_3, q_3\}$. Now consider how *C* classifies *D* using the nearest neighbor (NN) rule. It is clear from the distance relations that both $p_1$and q_1 are miss-classified. Therefore subset *C* is not *training-set consistent*. The counter-example in Fig. 2 can be easily extended to hold for any number of data points greater than six by adding additional points on both ends beyond p_1 with *y*-coordinate equal -1 and beyond q_1 with *y*-coordinate equal +1. Furthermore, in this way the resulting probability of error in classifying *D* with *C* can be as high as desired.

4. Concluding Remarks

Training-set consistency not with standing, Tomek's experiments demonstrated that in practice the modification of CNN may work better than CNN because fewer data points were kept and they more closely approximated the boundary of the NN-rule. However, the computational complexity of his algorithm is considerable. While he did not provide a complexity analysis, it is obvious from Fig. 1 that only the step for finding *C* (which is then fed into PROCEDURE CNN to take the place of *D*) already takes $O(n^3)$ time, where *n* is the size of *D*. For $n = 100{,}000$ as in OCR applications, this is not feasible. Furthermore, Tomek was not aware that he had re-discovered, in the definition of subset *C*, the idea of *Gabriel neighbors* well known in the biological sciences [GS69]. Given a set of points *D*, a pair(x, *y*) in *D* is called a *Gabriel* pair if all the remaining points in *D* lie outside the circle whose diameter is determined by (x, *y*). The reader can easily verify that this definition is equivalent to that in Fig. 1. In addition, if a Gabriel pair is connected by an edge, then the (now well known) Gabriel graph is obtained [MS80], [JT92] for which an $O(n \log n)$ time algorithm in the plane is now known and for which efficient expected time algorithms now exist in all dimensions using Voronoi diagrams [AB83], [Au91] or heuristics [TBP84]. The difference between the (complete) Gabriel graph of *D* and that computed in Fig. 1 is that in the complete Gabriel graph all pairs in *D* are candidates whereas in Fig. 1 only pairs that belong to different classes are included in *C*. More recently, a condensing scheme based on computing the complete Gabriel graph of *G* was proposed in [TBP84] where data points that have the property that all their Gabriel neighbors belong to one and the same class are then deleted from *D* to yield the final condensed set *E*. Note that this procedure is equivalent to using the initial subset *C* defined by Tomek as the *final* condensed set *E*, and doing away with CNN altogether. Whereas the algorithm of Tomek runs in at least $O(n^3)$ time, the algorithm in [TBP84] runs in time closer to $O(n^2)$. For more information concerning the efficient computation of other proximity graphs such as the relative neighborhood graph the reader is referred to [JT92].

5. References

[AB83] Avis, D. and Bhattacharya, B.K., "Algorithms for computing d-dimensional Voronoi diagrams and their duals," in *Computational Geometry*, Ed., F.P. Preparata, JAI Press, 1983, pp. 159-180.

[Au91] Aurenhammer, F., "Voronoi diagrams - A survey of a fundamental geometric data structure," *ACM Computing Surveys*, vol. 23, No. 3, September 1991, pp. 345-405.

[Ch74] Chang, C.-L., "Finding prototypes for nearest neighbor classifiers, *IEEE Trans. Computers*, vol. C-23, November 1974, pp.1179-1184.

[CH67] Cover, T. M. and Hart, P.E., "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. IT-13, No.1, 1967, pp. 21-27.

[De81] Devroye, L.P., "On the inequality of Cover and Hart in nearest neighbor discrimination," *IEEE Transactions on Pattern analysis and Machine Intelligence*, vol. PAMI-3, January 1981, pp. 75-78.

[DK82] Devijver, P. A. and Kittler, J., *Pattern Recognition: A Statistical Approach*, Prentice-

Hall, 1982.

[FM84] Fukunaga, K. and Mantock, J.M., “Nonparametric data reduction,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. PAMI-6, January 1984, pp. 115-118.

[FP70] Fisher, F.P. and Patrick, E.A., “A preprocessing algorithm for nearest neighbor decision rules,” *Proc. National Electronics Conf.*, Dec. 1970, pp. 481-485.

[Ga72] Gates, G.W., “The reduced nearest neighbor rule,” *IEEE Trans. Information Theory*, vol.IT-18, May 1972, pp. 431-433.

[GK79] Gowda, K.C. and Krishna, G., “The condensed nearest-neighbor rule using the concept of mutual nearest neighborhood,” *IEEE Transactions on Information Theory*, vol. IT-25, No.4, 1979, pp. 488-490.

[GS69] Gabriel, K. R. and Sokal, R. R., “A new statistical approach to geographic variation analysis,” *Systematic Zoology*, vol. 18, 1969, pp. 259-278.

[Ha68] Hart, P.E., “The condensed nearest-neighbor rule,” *IEEE Transactions on Information Theory*, vol. IT-4, May 1968, pp. 515-516.

[IIM85] Imai, H., Iri, M. and Murota, K., “Voronoi diagram in the Laguerre geometry and its applications,” *SIAM Journal of Computing*, vol. 14, No. 1, pp. 93-105.

[JT92] Jaromczyk, J. W. and Toussaint, G. T., “Relative neighborhood graphs and their relatives,” *Proc. IEEE*, vol. 80, No. 9, September 1992, pp. 1502-1517.

[KL87] Klein, R., *Concrete and Abstract Voronoi Diagrams*, Springer-Verlag, Berlin, 1987.

[MS80] Matula, D.W. and Sokal, R.R., “Properties of Gabriel graphs relevant to geographic variation research and the clustering of points in the plane,” *Geographical Analysis,* vol. 12, 1980, pp. 205-222.

[Ri75] Ritter, G.L., et al., “An algorithm for a selective nearest neighbor decision rule,” *IEEE Trans. Information Theory*, vol. IT-21, Nov. 1975, pp. 665-669.

[St77] Stone, C.J., “Consistent nonparametric regression,” *Annals of Statistics*, vol. 5, 1977 pp. 595-645.

[Sw72] Swonger, C.W., “Sample set condensation for a condensed nearest neighbor decision rule for pattern recognition,” in *Frontiers in Pattern Recognition*, Ed., S. Watanabe, Academic Press, 1972, pp. 511-526.

[TBP84] Toussaint, G. T., Bhattacharya, B. K. and Poulsen, R. S., “The application of Voronoi diagrams to non-parametric decision rules, “*Proc. Computer Science & Statistics: 16th*

Symposium of the Interface, Atlanta, Georgia, 1984.

[To76a] Tomek, I., “Two modifications of CNN,” *IEEE Trans. Systems, Man and Cybernetics*, vol. SMC-6, Nov. 1976, pp. 769-772.

[To76b] Tomek, I., “An experiment with the edited nearest neighbor rule,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-6, June 1976, pp. 448-452.

[Ul74] Ullmann, V.R., “Automatic selection of reference data for use in a nearest-neighbor method of pattern classification,” *IEEE Trans. Information Theory*, vol. IT-20, July 1974, pp. 541-544.

[Wi92] Wilfong, G., “Nearest neighbor problems,” Technical Report, AT&T Bell Laboratories, Murray Hill, New Jersey, 1992.