Proximity Graphs for Nearest Neighbor Decision Rules: Recent Progress

Godfried Toussaint *
School of Computer Science
McGill University
Montréal, Québec, Canada

Proceedings of INTERFACE-2002, *34th Symposium on Computing and Statistics* (theme: Geoscience and Remote Sensing), Ritz-Carlton Hotel, Montreal, Canada, April 17-20, 2002.

Abstract

In the typical nonparametric approach to pattern classification, random data (the training set of patterns) are collected and used to design a decision rule (classifier). One of the most well known such rules is the k-nearest-neighbor decision rule (also known as *instance-based learning*, and *lazy learning*) in which an unknown pattern is classified into the majority class among its k nearest neighbors in the training set. Several questions related to this rule have received considerable attention over the years. Such questions include the following. How can the storage of the training set be reduced without degrading the performance of the decision rule? How should the reduced training set be selected to represent the different classes? How large should k be? How should the value of k be chosen? Should all k neighbors be equally weighted when used to decide the class of an unknown pattern? If not, how should the weights be chosen? Should all the features (attributes) we weighted equally and if not how should the feature weights be chosen? What distance metric should be used? How can the rule be made robust to overlapping classes or noise present in the training data? How can the rule be made invariant to scaling of the measurements? Geometric proximity graphs such as Voronoi diagrams and their many relatives provide elegant solutions to most of these problems. After a brief and non-exhaustive review of some of the classical canonical approaches to solving these problems, the methods that use proximity graphs are discussed, some new observations are made, and avenues for further research are proposed.

1 Nearest-Neighbor Decision Rules

In the typical non-parametric classification problem (see Aha [2], Devroye, Gyorfy and Lugosi [37], Duda and Hart [38], Duda, Hart and Stork [39], McLachlan [70], O'Rourke and Toussaint [77]) we have available a set of d measurements or observations (also called a feature vector) taken from each member of a data set of n objects (patterns) denoted by $\{X, Y\} = \{(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)\}$, where X_i and Y_i denote, respectively, the feature vector on the ith object and the class label of that object. One of the most attractive decision procedures, conceived by

*This research was supported by NSERC and FCAR. e-mail: godfried@cs.mcgill.ca

Fix and Hodges in 1951, is the nearest-neighbor rule (1-*NN*-rule) [43]. Let Z be a new pattern (feature vector) to be classified and let X_j be the feature vector in $\{X,Y\} = \{(X_1,Y_1),(X_2,Y_2),...,(X_n,Y_n)\}$ closest to Z. The nearest neighbor decision rule classifies the unknown pattern Z into class Y_j.

A key feature of this decision rule (also called *lazy learning* [2], *instance-based learning* [3], and *memory-based reasoning* [100]) is that it performs remarkably well considering that no explicit knowledge of the underlying distributions of the data is used. Consider for example the two class problem and denote the *a priori* probabilities of the two classes by $P(C_1)$ and $P(C_2)$, the *a posteriori* probabilities by $P(C_1|X)$ and $P(C_2|X)$, and the mixture probability density function by

$$p(X) = P(C_1)p(X|C_1) + P(C_2)p(X|C_2),$$

where $p(X|C_i)$ is the class-conditional probability density function given class $C_i, i = 1,2$. In 1967 Cover and Hart [21] showed, under some continuity assumptions on the underlying distributions, that the asymptotic error rate of the 1-*NN* rule, denoted by P_e[1-*NN*] is given by

$$P_e[1-NN] = 2\mathbf{E}_X[P(C_1|X)P(C_2|X)],$$

where $\mathbf{E}_X$ denotes the expected value with respect to the mixture probability density function $p(X)$. They also showed that P_e[1-*NN*] is bounded from above by twice the Bayes error (the error of the best possible rule). More precisely, and for the more general case of M pattern classes the bounds proved by Cover and Hart [21] are given by:

$$P_e \leq P_e[1-NN] \leq P_e(2 - MP_e/(M-1)),$$

where P_e is the optimal Bayes probability of error.

Stone [101] and Devroye [36] generalized these results by proving the bounds for all distributions. In other words, the nearest neighbor of Z contains at least half of the total discrimination information contained in an infinite-size training set. Furthermore, a simple generalization of this rule called the k-*NN*-rule, in which a new pattern Z is classified into the class with the most members present among the k nearest neighbors of Z in $\{X,Y\}$, can be used to obtain good estimates of the Bayes error (Fukunaga and Hostetler [46]) and its probability of error asymptotically approaches the Bayes error (Devroye et al. [37]).

The measure P_e[1-*NN*] turns up in a surprising variety of related problems sometimes in disguise. For example, it is also the error rate of the *proportional prediction* randomized decision rule considered by Goodman and Kruskal [48] (see also Toussaint [111]). Devijver and Kittler [33] and Vajda [126] refer to it as the *quadratic entropy.* Mathai and Rathie [68] call it the *harmonic mean coefficient.* It is also closely related to the *Bayesian distance* (Devijver [32]) and the *quadratic mutual information* (Toussaint [108]). Incidentally, the *Bayesian distance* is called the *cross-category feature importance* in the instance-based learning literature (Stanfill and Waltz [100], Creecy et al.[22]). Furthermore, it is identical to the asymptotic probability of correct classification of the 1-*NN*-rule given by P_c[1-*NN*] = 1 - P_e[1-*NN*]. The error probability P_e[1-*NN*] also shares a property with Shannon's measure of equivocation. Both are special cases of the equivocation of order β (Toussaint [109], [113]).

There is a vast literature on the subject of nearest neighbor classification which will not be reviewed here. The interested reader is referred to the comprehensive treatment by Devroye, Györfi and Lugosi [37] and the collected papers in the 1991 volume edited by Dasarathy [24]. For more on the information measures closely related to the measure P_e[1-*NN*] the reader is referred to Mathai and Rathie [68] (see also Toussaint [110]).

In the past many pattern recognition practicioners have unfairly criticized the NN-rule on the grounds of the mistaken assumptions that (1) *all* the data $\{X, Y\}$ must be stored in order to implement such a rule, (2) to determine the nearest neighbor of a pattern to be classified, distances must be computed between the unknown vector Z and *all* members of $\{X, Y\}$, and (3) nearest neighbor rules are not well suited for fast parallel computation. As we shall see below, all three of these assumptions are incorrect and computational geometric progress in the 1980's and 1990's along with faster and cheaper hardware has made the *k-NN*-rules a practical reality for pattern recognition applications in the 21st Century.

2 Reducing the Size of the Stored Training Data

2.1 Hart's condensed rule and its relatives

In 1968 Hart was the first to propose an algorithm for reducing the size of the stored data for the nearest neighbor decision rule [51]. Hart defined a *consistent* subset of the data as one that classified the remaining data correctly with the nearest neighbor rule. He then proposed an algorithm for selecting a consistent subset by heuristically searching for data that were near the decision boundary. The algorithm is very simple. Let C denote the desired final consistent subset. Initially C is empty. First a random element from $\{X, Y\}$ is transferred to C. Then C is used as a classifier with the *1-NN* rule to classify all the remaining data in $\{X, Y\}$. During this scan of $\{X, Y\}$ whenever an element is incorrectly classified by C it is transferred from $\{X, Y\}$ to C. Thus $\{X, Y\}$ is shrinking and C is growing. This scan of $\{X, Y\}$ is repeated as long as least one element is transferred from $\{X, Y\}$ to C during a complete pass of the remaining data in $\{X, Y\}$. The goal of the algorithm is to keep only a *subset* of the data $\{X, Y\}$ that are necessary to determine the decision boundary of all the data $\{X, Y\}$. The motivation for this heuristic is the intuition that data far from the decision boundary are not needed and that if an element is misclassified it must lie close the the decision boundary. By construction the resulting reduced set C classifies all the training data $\{X, Y\}$ correctly and hence it is referred to here as a *training-set consistent* subset. In the literature Hart's algorithm is called *CNN* and the resulting subset of $\{X, Y\}$ is called a *consistent* subset. Here the longer term *training-set consistent* is used in order to distingish it from another interesting type of subset: one that determines *exactly* the same decision boundary as the entire training set $\{X, Y\}$. The latter kind of subset will be called *decision-boundary consistent.* Clearly decision-boundary consistency implies training-set consistency but the converse is not necessarily true. Empirical results have shown that Hart's *CNN* rule considerably reduces the size of the training set and does not greatly degrade performance on a separate testing (validation) set. It is also easy to see that using a naive brute-force algorithm the complexity of computing the condensed subset of $\{X, Y\}$ is $O(n^3)$. However, the method does not in general yield a minimal-size consistent subset and unfortunately may change the decision boundary of the original training set. Recently several theoretical results on *CNN* have been obtained by Devroye et al. [37].

In 1987 Kibler and Aha [59] proposed an algorithm called the *growth-additive* algorithm which consists of only one pass of Hart's *CNN* rule. Such an algorithm is of course not training-set consistent. On the other hand a naive implementation of it runs in $O(n^2)$ worst-case time.

CNN may keep points far from the decision boundary. To combat this Gates [47] proposed what he called the *reduced nearest neighbor rule* or *RNN*. *RNN* consists of first performing *CNN* and then adding a post-processing step. In this post-processing step elements of C are visited and deleted from C if their deletion does not result in misclassifying any elements in $\{X, Y\}$. Experimental results confirmed that *RNN* yields a slightly smaller training-set consistent subset of $\{X, Y\}$ than that obtained with *CNN* [47].

Tomek [105] proposed a modification of *CNN* in which a preliminary pass of $\{X, Y\}$ is made to select an order-independent special subset of $\{X, Y\}$ that lies close to the decision boundary. After this preprocessing step his method proceeds in the same manner as *CNN* but instead of processing $\{X, Y\}$ it works on the special subset so preselected. The algorithm to preselect the special subset of $\{X, Y\}$ consists of keeping all pairs of points (X_i, Y_i), (X_j, Y_j) such that $Y_i \neq Y_j$ (the two points belong to different classes) and the *diametral* sphere determined by X_i and X_j does not contain any points of $\{X, Y\}$ in its interior. Such pairs are often called *Tomek links* in the literature. Clearly, pairs of points far from the decision boundary will tend to have other points in the interior of their diametral sphere. It is claimed in [105] that the resulting subset of $\{X, Y\}$ is training-set consistent. However, Toussaint [120] demonstrated a counter-example. It should be noted that Tomek's preselected non-consistent subset using the diametral sphere test *implicitly* computes a subgraph of the Gabriel graph [58] of $\{X, Y\}$, a graph admirably suited for condensing the training data that will be discussed later.

2.2 Order-independent subsets

CNN, *RNN* and Tomek's modification of *CNN* all have the undesirable property that the resulting reduced consistent subsets are a function of the order in which the data are processed. Several attempts have been made to obtain training-set consistent subsets that are less sensitive to the order of presentation of the data. One class of methods with this goal suggested by Alpaydin [4] applies the above methods several times (processing the data in a different random order each time) to obtain a group of training-set consistent subsets. Then a voting technique among these groups is used to make the final decision.

A successful solution to obtaining order-independent training-set consistent subsets by generalizing Hart's *CNN* procedure was proposed by Devi and Murty [31]. Recall that in Hart's procedure the subset C starts with a single random element from $\{X, Y\}$ and subsequently each time an element from $\{X, Y\}$ is misclassified it is transferred to C. In other words transfers are made one at a time and class-membership is not an issue. In contrast, the method of Devi and Murty [31], which they call the modified condensed nearest neighbor rule (*MCNN*) initializes the reduced set (call it MC) by transferring, in batch mode, one representative of each class from $\{X, Y\}$ to MC. Subsequently MC is used to classify *all* elements of $\{X, Y\}$. Then from each class of the resulting misclassified patterns a representative is transferred to MC (again in batch mode). This process is repeated until all the patterns in $\{X, Y\}$ are classified correctly. Note that if at some stage there is a class, say C_i, that has no misclassified patterns using MC, then no representative is transferred from $\{X, Y\}$ to MC at that stage. Hence the most difficult classes (the last ones to be completely correctly classified) receive more representatives in MC. Thus this approach provides a natural way to automatically decide how many representatives each class should be alloted and how they should be distributed.

2.3 Minimal size training-set consistent subsets

The first researchers to deal with computing a *minimal-size* training-set consistent subset were Ritter et al. [89]. They proposed a procedure they called a *selective* nearest neighbor rule *SNN* to obtain a minimal-size training-set consistent subset of $\{X, Y\}$, call it S, with one additional property that Hart's *CNN* does not have. Any training-set consistent subset C obtained by *CNN* has the property that every element of $\{X, Y\}$ is nearer to an element in C of the same class than to any element in C of a different class. On the other hand, the training-set consistent subset S of Ritter et al. [89] has the additional property that every element of $\{X, Y\}$ is nearer to an element in S of the same class than to any element, in the *complete* set, $\{X, Y\}$ of a different class. This additional property of *SNN* tends to keep points closer to the decision boundary than does *CNN*. The additional property allows Ritter et al. [89] to compute the selected subset S without testing all possible subsets of $\{X, Y\}$. Nevertheless, their algorithm still runs in time exponential in n (see Wilgong [131]) in the worst case. However, Wilson and Martinez [133] and Wilson [134] claim that the average running time of *SNN* is $O(n^3)$. Furthermore, experimental results indicate that the resulting cardinality of S is about the same as that of the *reduced* nearest neighbor consistent subsets of Gates [47]. Hence the heavy computational burden of *SNN* does not make it competitive with *RNN*.

In 1994 Dasarathy [25] proposed a complicated algorithm intended to compute a *minimal-size* training-set consistent subset but did not provide a proof of optimality. The algorithm uses a subset of $\{X, Y\}$ that he calls the Nearest Unlike Neighbor (*NUN*) subset [26]. Given an element X_i of $\{X, Y\}$, the element of $\{X, Y\}$ closest to X_i but belonging to a different class is called the nearest unlike neighbor of X_i. The *NUN* subset consists of all points in $\{X, Y\}$ that are nearest unlike neighbors of one or more elements of $\{X, Y\}$. The algorithm yields a consistent subset of $\{X, Y\}$ which he calls the *MCS* (Minimal Consistent Subset). Extensive experiments led him to conjecture that his algorithm generated an *MCS* that is minimal-size. However, counter-examples to this claim have been found by Kuncheva and Bezdek [64], Cerverón and Fuertes [15] and Zhang and Sun [136].

Wilson and Martinez [133] rediscovered the idea of using the nearest unlike neighbors to reduce the size of the training-set consistent subsets. They call the training-set condensing algorithms "instance pruning techniques" and refer to the nearest unlike neighbor as the *nearest enemy*. They also propose three algorithms for computing training-set consistent subsets. Many other similar algorithms can be found in the literature on *instance-based* and *lazy* learning (Mantaras and Armengol [29], Aha, Kibler and Albert [3] and Aha [1], [2]).

Wilfong [131] showed in 1991 that the problem of finding the smallest size training-set consistent subset is NP-complete when there are three or more classes. Furthermore, he showed that even for only two classes, finding the smallest size training-set consistent *selective* subset (Ritter et al. [89]) is also NP-complete.

2.4 Prototype generation methods

The techniques discussed above have in common that they select a subset of the training set as the final classifier. There exists also a class of techniques that do not have this restriction when searching for a good set of prototypes. These methods are sometimes called *prototype generation* methods (also *replacement* techniques). One of the first such algorithms, proposed in 1974 by Chang [16], repeatedly merges the two nearest neighbors of the same class as long as this merger does not increase the error rate on the training set. One drawback of Chang's method is that it may yield prototypes that do not characterize well the training set in terms of generalization. To combat this Mollineda et al. [73] modified Chang's algorithm to merge *clusters* (rather than pairs of data points) based on several geometric criteria. Thus the technique resembles hierarchical bottom-up clustering guided by a constraint on the resulting error rate on the training set. Bezdek et al. [9] proposed another modification of Chang's method and demonstrated that it produced a smaller set of prototypes for the well known Iris data set.

Salzberg et al. [91] proposed a novel and more extreme version of the above methods in which they focus on the desired decision boundary and ignore the training set. Instead they design a minimal size set of prototypes to realize the desired decision boundary by synthesizing *best-case* prototypes.

2.5 Optimization methods

There have been many approaches that use approximate optimization techniques to find a subset close to the smallest size training-set consistent subset. Such methods include tabu search (Cerverón and Ferri [14], Zhang and Sun [136]), gradient descent and deterministic annealing (Decaestecker [30]), genetic algorithms (Kuncheva [63], Kuncheva and Jain [62], Chang and Lippmann [17]), evolutionary learning (Zhao and Higuchi [138], Zhao [137]), bootstrapping (Saradhi and Murty [95]) and other random search techniques (Lipowezky [66]). Alternately, some techniques first select the prototype subset and subsequently minimize the error rate (Bermejo and Cabestany [8]). Liu and Nakagawa [67] recently compared 11 optimization methods to each other. Unfortunately they did not compare those techniques to the proximity-graph methods (Toussaint, Bhattacharya and Poulsen [122]) to be discussed below.

2.6 Decision-boundary generation methods

Consider two elements $\{X_i, Y_i\}$ and $\{X_j, Y_j\}$ in $\{X, Y\}$ such that $Y_i \neq Y_j$. If the two points are used in the 1-*NN* rule they implement a linear decision boundary which is the hyperplane that orthogonally bisects the line segment joining $\{X_i, Y_i\}$ and $\{X_j, Y_j\}$. Thus when a subset of $\{X, Y\}$ is being selected in the above methods the hyperplanes are being chosen implicitly. However, we could just as well be selecting these hyperplanes explicitly. When classifiers are designed by manipulating more hyperplanes than there are pattern classes they are called *piece-wise* linear classifiers (Sklansky and Michelotti [97]). There is a vast field devoted to this problem which is beyond the scope of this study and the interested reader is referred to the book by Nilsson [75]. One can also generate non-parametric decision boundaries with other surfaces besides hyperplanes. Priebe et al. [85], [84] model the decision surfaces with balls.

3 Editing to Improve Performance

Methods that have as their goal the improvement of recognition acuracy rather than data reduction are called editing rules. In 1972 Wilson [132] conceived the idea of editing and proposed the following algorithm.

PREPROCESSING
A. For each i:
  1. Find the k-nearest neighbors to X_i among $\{X, Y\}$ (not counting X_i).

2. Classify X_i to the class associated with the largest number of points among the k-nearest neighbors (breaking ties randomly).

B. Edit $\{X, Y\}$ by deleting all the points misclassified in the foregoing.

DECISION RULE

Classify a new unknown pattern Z using the 1-*NN* rule with the *edited* subset of $\{X, Y\}$.

This simple editing scheme is so powerful that the error rate of the 1-*NN* rule that uses the edited subset converges to the Bayes error. We remark here that a gap in the proof of Wilson [132] was pointed out by Devijver and Kittler [34] but alternate proofs were provided by Wagner [128] and Penrod and Wagner [82].

Wilson's deleted nearest neighbor rule deletes all the data misclassified by the k-*NN* majority rule. A modified editing scheme was proposed in 2000 by Hattori and Takahashi [53] in which the data X_i are kept only if *all* their k-nearest neighbors belong to the same class as that of X_i. Thus only the strongest correctly classified data are kept.

Tomek [104] and Devijver and Kittler [33] proposed the repeated application of Wilson editing until no points are discarded. Devijver and Kittler [33] showed that with this scheme, which they call *multi-edit*, repeated editing with the 1-*NN* will lead to the Bayes error rate.

4 Weighting the Neighbors

The k-*NN* rule makes a decision based on the majority class membership among the k nearest neighbors of an unknown pattern Z. In other words every member among the k nearest neighbors has an equal say in the vote. However, it is natural to give more weight to those members that are closer to Z. In 1966 Royall [90] suggested exactly such a scheme where the i-th neighbor receives weight w_i, $w_1 \geq w_2 \geq ... \geq w_k$ and $w_1 + w_2 + ... + w_k = 1$. In 1976 Dudani [40] proposed such a weighting scheme where the weight given to X_i is inversely proportional to the distance between Z and X_i. He also showed empirically that for small training sets and certain distributions this rule gave higher recognition acuracy than the standard k-*NN* rule. Similar results were obtained by Priebe [83] when he applied a *randomly* weighted k-*NN* rule to an *olfactory classification* problem. However, these results do not imply that the weighted rule is better than the standard rule asymptotically (Bailey and Jain [6], Devroye et al. [37]).

5 Weighting the Features

Considerable work has been done on trying to improve the nearest neighbor decision rules by weighting the features (attributes, measurements) differently. It is natural to try to put more weight on features that are better. Hence much effort has been directed at evaluating and comparing measures of the goodness of features. This problem is closely related to the vast field of feature selection where one is interested in selecting a subset of features with which to design a classifier. This is similar to setting the weights to "one" for the features selected and "zero" to the ones discarded.

One popular method for measuring goodness is with measures of information. For example, Lee and Shin [65] propose an enhanced nearest neighbor learning algorithm, that has applications to relational data bases, in which they use information theory to calculate the weights of the attributes. More specifically they use the Hellinger divergence, a measure equivalent to the Bhattacharya coefficient, the Matusita distance and the affinity (Matusita [69], McLachlan [70]) to calculate the weights automatically. This measure of distance between the class-conditional probability distributions, is closely related to the Bayes probability of error (Hellman and Raviv [54], Toussaint [112], [114], [115], Bhattacharya and Toussaint [11]).

Wettschereck and Aha [129] and Wettschereck et al. [130] compare several methods for weighting features in nearest neighbor rules and claim that the mutual information (Cover and Thomas [18]) gives good results. As with feature selection, one must be careful when calculating and evaluating weights of features independently of each other and then using them together (Toussaint [106]).

We close this section by mentioning a third popular measure for weighting features called the *cross-category feature-importance* measure by Wettschereck et al. [130], Wettschereck and Aha [129], Stanfill and Waltz [100], and Creecy et al. [22]. This measure is equivalent to the asymptotic probability of correct classification of the nearest neighbor decision rule and also has several other names such as the *Bayesian distance* (Devijver [32]) as mentioned in the introduction. While such weighting schemes may sometimes improve results in practice there are no guarantees. Most unsettling is the fact that even when using the Bayes probability of correct classification (the ultimate criterion) as an evaluation measure, and even when the features are independent in each and every pattern class, selecting the best individual features may actually result in obtaining the worst possible feature subset (Toussaint [107], Cover [19], Cover and Van Campenhout [20]).

6 Choice of Metric

There has been considerable effort spent on finding the "optimal" metric for use in the distance calculations. Empirical improvements in accuracy are often obtained when the metric adapts locally to the distribution of the data (Short and Fukunaga [96], Friedman [45], Hastie and Tibshirani [52], Ricci and Avesani [88]).

Several elegant nearest neighbor rules have been devised that are scale-invariant. One method suggested by Olshen [76] and Devroye [35] uses empirical distances defined in terms of order statistics along the d coordinate axes. Another technique suggested by Ichino and Sklansky [55] and Devroye et al. [37] is the *rectangle-of-influence* graph decision rule. An unknown pattern Z is classified by a majority rule among the rectangle-of-influence neighbors of Z in $\{X, Y\}$. A point X_i in $\{X, Y\}$ is such a neighbor if the smallest axis-parallel hyper-rectangle containing both Z and X_i contains no other points of $\{X, Y\}$. Devroye et al. [37] call this rule the *layered nearest neighbor rule* and have shown that if there are no ties it is asymptotically Bayes optimal.

7 Proximity Graph Methods

7.1 Proximity graphs

The most natural proximity graph defined on a set of points $\{X, Y\}$ is the *nearest neighbor graph* or *NNG*. Here each point in $\{X, Y\}$ is joined by an edge to its nearest neighbor (Paterson and Yao [81]). Another well known proximity graph is the minimum spannig tree (*MST*) Zahn [135]. In 1980 the relative neighborhood graph (*RNG*) was proposed as a tool for extracting the shape of a planar pattern (see Toussaint [118], [116], [123]). However, such definitions are readily extended to higher dimensions. For computing the RNG in d dimensions see Su and Chang [102]. Proximity graphs have many applications in pattern recognition (see Toussaint [117], [124], [122]). There is a vast literature on proximity graphs and it will not be reviewed here. The reader is directed to Jaromczyk and Toussaint [58] for a start. The most well known proximity graphs besides those mentioned above are the Gabriel graph GG and the Delaunay triangulation DT. All these are nested together in the following relationship:

$$NNG \subseteq MST \subseteq RNG \subseteq GG \subseteq DT \tag{1}$$

7.2 Decision-boundary-consistent subsets

In 1978 Dasarathy and White were the first to characterize and compute explicitly the decision surfaces of nearest neighbor rules [23] but only for the case of $d = 2, 3$.

In 1979 Toussaint and Poulsen [124] were the first to use d-dimensional Voronoi diagrams to delete "redundant" members of $\{X, Y\}$ in order to obtain a subset of $\{X, Y\}$ that implements *exactly* the same decision boundary as would be obtained using all of $\{X, Y\}$. For this reason the method is called *Voronoi condensing*. The algorithm in [124] is very simple. Two points in $\{X, Y\}$ are called *Voronoi neighbors* if their corresponding Voronoi polyhedra share a face. First mark each point X_i if all its Voronoi neighbors belong to the same class as X_i. Then discard all marked points. The remaining points form the Voronoi condensed subset $\{X, Y\}$. Voronoi condensing does not change the error rate of the resulting decision rule because the nearest nighbor decision boundary with the reduced set is identical to that obtained by using the entire set. For this reason the Voronoi condensed subset is called *decision-boundary consistent*. Clearly decision-boundary consistency implies training-set consistency but the converse is not necessarily so. The most important consequence of this property is that all the theory developed for the 1-*NN* rule continues to hold true when the rule is preprocessed with Voronoi condensing.

Sixteen years later Murphy, Brooks and Kite [74] rediscovered the above algorithm in the context of neural network design and called it *network reduction*. Unaware of the above references in 1999 Esat [42] again rediscovered Voronoi condensing and called it *Voronoi polygon reduction*. It should be noted that the neural networks for nearest neighbor classification proposed in [74] and [42] are complicated and much simpler networks are possible (see Toussaint [121]).

In 1998 Bhattacharya and Kaller [10] extended the above methods to the k-nearest neighbor rules. They call the decision-boundary consistent condensing *exact thinning* and otherwise *inexact thinning*. They proposed a proximity graph they call the *k-Delaunay* graph and showed how exact thinning can be performed with this graph.

7.3 Condensing prototypes via proximity graphs

In 1985 Toussaint, Bhattacharya and Poulsen [122] generalized Voronoi condensing so that it would discard more points in a judicious and organized manner so as not to degrade performance unnecessarily. The dual of the Voronoi diagram is the Delaunay triangulation. In this setting Voronoi condensing can be described as follows. Compute the Delaunay triangulation of $\{X, Y\}$. Mark a vertex X_i of the triangulation if all its (graph) neighbors belong to the same class as that of X_i. Finally discard all the marked vertices. The remaining points of $\{X, Y\}$ form the Voronoi condensed set. The methods proposed in [122] substitute the Delaunay triangulation by a subgraph of the triangulation. Since a subgraph has fewer edges, its vertices have lower degree on the average. This means the probability that all the graph neighbors of X_i belong to the same class as that of X_i is higher, which implies more elements of $\{X, Y\}$ will be discarded. By selecting an appropriate subgraph of the Delaunay triangulation one can control the number of elements of $\{X, Y\}$ that are discarded. Furthermore by virtue of the fact that the graph is a subgraph of the Delaunay triangulation and that the latter yields a decision-boundary consistent subset, we are confident in degrading the performance as little as possible. Experimental results obtained in [122] suggested that the Gabriel graph is the best in this respect.

Also in 1985 and independently of Toussaint, Bhattacharya and Poulsen [122], Ichino and Sklansky [55] suggested the same idea but with a different proximity graph that is not necessarily a subgraph of the Delaunay triangulation. They proposed a graph which they call the *rectangular-influence* graph or RIG defined as follows. Two points X_i and X_j in $\{X, Y\}$ are joined by an edge if the smallest axis-parallel hyper-rectangle that contains both X_i and X_j contains no other point of $\{X, Y\}$. Not surprisingly, condensing the training set with the RIG does not guarantee a decision-boundary consistent subset. On the other hand recall that the RIG has the nice property that it is scale-invariant which can be very useful in classification problems.

In 1998 Bhattacharya and Kaller [10] proposed a proximity graph they call the *k-Gabriel* graph and show how inexact thinning can be performed with this graph. The *k-Gabriel* graph is much easier to compute than the *k-Delaunay* graph and yields good results.

7.4 Editing via proximity graphs

Sánchez, Pla and Ferri [93] extended Wilson's [132] editing idea to incorporate proximity graphs. Their algorithms mimic Wilson's algorithm. Instead of discarding points correctly classified by the k-nearest neighbor rule they are discarded if they are correctly classified with the graph neighbors. They empirically investigated the relative neighborhood graph and the Gabriel graph and found that Gabriel graph editing was the best.

7.5 Combined editing and condensing via proximity graphs

Editing by itself smooths the decision boundary and improves performance with finite sample size. However, it tends not to discard much data. Therefore to reduce the size of the training set condensing is necessary. There has been much research lately at exploring the synergy between editing and condensing techniques (Dasarathy and Sánchez [27], Dasarathy, Sánchez and Townsend [28]). Sánchez, Pla and Ferri [93] also explored the interaction between editing and condensing. One conclusion of these studies is that editing should be done before condensing to obtain the best results. An extensive experimental comparison of 26 techniques shows that the best approach in terms of both recognition acuracy and data compression is to first edit with either the Gabriel graph or the relative neighborhood graph and subsequently condense with the minimal-consistent subset (*MCS*) algorithm (Dasarathy, Sánchez and Townsend [28]).

7.6 Piece-wise classifier design via proximity graphs

Proximity graphs have also found use in designing piece-wise linear and spherical classifiers. Sklansky and Michelotti [97] and Park and Sklansky [79], [80] use the

Gabriel graph edges that connect points of *different* classes (which they call Tomek links) to guide the selection of the final hyperplanes used to define the decision boundary. In particular they require the selection of hyperplanes that intersect all such edges. More recently Tenmoto, Kudo and Shimbo [103] use the Gabriel graph edges between classes (Tomek links) only as a starting point for the initial position of the hyperplanes and subsequently apply error-correction techniques to change the position of the hyperplanes if the local performance improves.

As mentioned earlier, one can also generate non-parametric decision boundaries with other surfaces besides hyperplanes. Priebe et al. [85], [84] model the decision surfaces with balls and use proximity graphs called *catch digraphs* to determine the number, location and size of the balls.

7.7 Cluster analysis and validation via proximity graphs

One of the most natural ways to select prototypes to represent a class in pattern recognition is to perform a cluster analysis on the training data of the class in question (Duda et al. [39]). The number and shape of the resulting clusters can then guide the designer in selecting the prototypes. Obviously one can bring the entire available clustering and vector quantization arsenals to bear down on this problem (Jardine and Sibson [57], Jain and Dubes [56], Kohonen [61], Baras and Subhrakanti [7]). However, the most powerful and robust methods for clustering turn out to be those based on proximity graphs. Florek et al. [44] were the first to propose the minimum spanning tree proximity graph as a tool in classification. The minimum spanning tree contains the nearest-neighbor graph as a subgraph. Zahn [135] demosntrated the power and versatility of the minimum spanning tree when applied to many pattern recognition problems. These techniques were later generalized by using other proximity graphs by Urquhart [125]. Another generalization of the nearest-neighbor graph is the k-nearest-neighbor graph. This graph is obtained by joining each point with an edge to its k nearest neighbors. Brito te al. [13] study the connectivity of the k-nearest-neighbor graph and apply it to clustering and outlier detection. Once a clustering is obtained it is desirable to perform a cluster-validation test. Pal and Biswas [78] propose some new indices of cluster validity based on three proximity graphs: the minimum spanning tree, the relative neighborhood graph and the Gabriel graph, and they show that for an interesting class of problems they outperform the existing indices.

7.8 Proximity-graph-neighbor decision rules

The classical approaches to *k-NN* decision rules are rigid in at least two ways: (1) they obtain the k nearest neighbors of the unknown pattern Z based purely on distance information, and (2) the parameter k is fixed. Thus they disregard how the nearest neighbors are distributed around Z. In 1985 Ichino and Sklansky [55] proposed the rectangle-of-influence graph neighbor rule in which Z is classified by a majority vote of its rectangle-of-influence graph neighbors. Recently new geometric definitions of neighborhoods have been proposed and new nearest neighbor decision rules based on other proximity graphs (Jaromczyk and Toussaint [58]) have been investigated. Devroye et al. [37] proposed the *Gabriel neighbor rule* which takes a majority vote among all the Gabriel neighbors of Z among $\{X, Y\}$. Sánchez, Pla and Ferri [92], [94] proposed similar rules with other graphs as well as the Gabriel and relative neighborhood graphs. Thus both the value of k and the distance of the neighbors vary locally and adapt naturally to the distribution of the data around Z. Note that these methods also automatically and implicitly assign different "weights" to the nearest geometric neighbors of Z.

8 Open Problems and New Directions

In 1985 Kirkpatrick and Radke [60] (see also Radke [86]) proposed a generalization of the Gabriel and relative neighborhood graphs which they called β-skeletons, where β is a parameter that determines the shape of the neighborhood of two points that must be empty of other points before the two points are joined by an edge in the graph. It is possible that for a suitable value of β these proximity graphs may yield better training set condensing results than the those obtained with the Gabriel graph. This should be checked out experimentally. It should be noted that there is a close relationship between β-skeletons and the concept of *mutual nearest neighbors* used by Gowda and Krishna [49]. For $\beta = 1$ an edge in the β-skeleton has the property that the two points it joins are the mutual nearest neighbors of each other. For further references on computing β-skeletons the reader is referred to the paper by Rao and Mukhophadhyay [87].

In the early 1980's a graph named the *sphere-of-influence* graph was proposed that was originally intended to capture the low-level perceptual structure of visual scenes consisting of dot-patterns, or more precisely, a set of n points in the plane (see Toussaint [119]). It was also conjectured that (although not planar) this graph had a linear number of edges. Avis and Horton [5] showed in 1985 that the number of edges in the sphere-of-influence graph of n points was bounded above by $29n$. The best upper bound until recently remained fixed at $17.5n$. Finally in 1999 Michael Soss [99] brought this bound down to $15n$. Avis has conjectured that the correct upper bound is $9n$ and has found examples that require $9n$ edges, so the problem is still open. More relevant to the topic of interest here is the fact that the sphere-of-influence graph yields a natural clustering of points completely automatically without the need of tuning any parameters. Furthermore, Guibas, Pach and Sharir showed that even in higher dimensions it has $O(n)$ edges for fixed dimension [50]. Soss has also given results on the number of edges for metrics other than Euclidean [98]. Finally, Dwyer [41] has some results on the expected number of edges in the sphere-of-influence graph. For two recent papers with many references to recent results on sphere-of-influence graphs the reader is referred to Michael and Quint [72] and Boyer, et al. [12]. To date the sphere-of-influence graph has not been explored for applications to nearest neighbor decision rules. Even more recently several new classes of proximity graphs have surfaced. These include the sphere-of-attraction graphs of McMorris and Wang [71], and the class-cover catch digraphs of De Vinney and Priebe [127]. It would be interesting to compare all these graphs on the problems discussed in this paper.

Recall that Gordon Wilfong [131] showed in 1991 that the problem of finding the smallest size training-set consistent subset is NP-complete when there are more than two classes. The complexity for the case of *two* classes remains an open problem.

References

[1] D. W. Aha. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. *International Journal of Man-Machine Studies*, 36:267–287, 1992.

[2] D. W. Aha, editor. *Lazy Learning.* Kluwer, Norwell, MA, 1997.

[3] D. W. Aha, D. Kibler, and M. Albert. Instance-based learning algorithms. In *Machine Learning, 6*, pages 37–66. Kluwer, Boston, 1991.

[4] Ethem Alpaydin. Voting over multiple condensed nearest neighbors. *Artificial Intelligence Review*, 11:115–132, 1997.

[5] David Avis and Joe Horton. Remarks on the sphere of influence graphs. *Annals of the New York Academy of Sciences*, 440:323–327, 1982.

[6] T. Bailey and A. Jain. A note on distance-weighted k-nearest neighbor rules. *IEEE Transactions on Systems, Man, and Cybernetics*, 8:311–313, 1978.

[7] John S. Baras and Subhrakanti Dey. Combined compression and classification with learning vector quantization. *IEEE Transactions on Information Theory*, 45:1911–1920, 1999.

[8] Sergio Bermejo and Joan Cabestany. Adaptive soft k-nearest-neighbor classifiers. *Pattern Recognition*, 32:2077–2979, 1999.

[9] James C. Bezdek, Thomas R. Reichherzer, Gek Sok Lim, and Yianni Attikiouzel. Multiple-prototype classifier design. *IEEE Trans. Systems, Man and Cybernetics - Part C: Applications and Reviews*, 28:67–79, 1998.

[10] Binay Bhattacharya and Damon Kaller. Reference set thinning for the k-nearest neighbor decision rule. In *Proceedings of the 14th International Conference on Pattern Recognition*, volume 1, 1998.

[11] Binay K. Bhattacharya and Godfried T. Toussaint. An upper bound on the probability of misclassification in terms of Matusita's measure of affinity. *Annals of the Institute of Statistical Mathematics*, 34:161–165, 1982.

[12] E. Boyer, L. Lister, and B. Shader. Sphere of influence graphs using the sup-norm. *Mathematical and Computer Modelling*, 32:1071–1082, 1999.

[13] M. R. Brito, E. L. Cháves, A. J. Quiroz, and J. E. Yukich. Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection. *Statistics and Probability Letters*, 35:33–42, 1997.

[14] Vicente Cerverón and Francesc J. Ferri. Another move toward the minimum consistent subset: a tabu search approach to the condensed nearest neighbor rule. *IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics*, 31:408–413, 2001.

[15] Vicente Cerverón and A. Fuertes. Parallel random search and Tabu search for the minimum consistent subset selection problem. In *Lecture Notes in Computer Science*, pages 248–259. Springer, Berlin, 1998.

[16] C. L. Chang. Finding prototypes for nearest neighbor classifiers. *IEEE Transactions on Computers*, 23:1179–1184, 1974.

[17] I. E. Chang and R. P. Lippmann. Using genetic algorithms to improve pattern classification performance. *Advances in Neural Information Processing*, 3:797–803, 1991.

[18] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, New York, 1991.

[19] Thomas M. Cover. The best two independent measurements are not the two best. *IEEE Trans. Systems, Man, and Cybernetics*, 4:116–117, 1974.

[20] Thomas M. Cover and Jan Van Campenhout. On the possible orderings in the measurement selection problem. *IEEE Trans. Systems, Man, and Cybernetics*, 7:657–661, 1977.

[21] Thomas M. Cover and Peter E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.

[22] Robert H. Creecy, Brij M. Masand, Stephen J. Smith, and David L. Waltz. Trading MIPS and memory for knowledge engineering. *Communications of the ACM*, 35:48–63, August 1992.

[23] Balakrishnan Dasarathy and Lee J. White. A characterization of nearest-neighbor rule decision surfaces and a new approach to generate them. *Pattern Recognition*, 10:41–46, 1978.

[24] Belur V. Dasarathy, editor. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques.* IEEE Comp. Society Press, Los Alamitos, 1991.

[25] Belur V. Dasarathy. Minimal consistent set (MCS) identification for optimal nearest neighbor decision system design. *IEEE Trans. on Systems, Man and Cybernetics*, 24:511–517, 1994.

[26] Belur V. Dasarathy. Nearest unlike neighbor (NUN): an aid to decision confidence estimation. *Optical Engineering*, 34:2785–2792, 1995.

[27] Belur V. Dasarathy and J. S. Sánchez. Tandem fusion of nearest neighbor editing and condensing algorithms - data dimensionality effects. In *Proc. 15th Int. Conf. on Pattern Recognition*, pages 692–695, September 2000.

[28] Belur V. Dasarathy, J. S. Sánchez, and S. Townsend. Nearest neighbor editing and condensing tools - synergy exploitation. *Pattern Analysis and Applications*, 3:19–30, 2000.

[29] Ramon Lopez de Mantaras and Eva Armengol. Inductive and lazy methods. *Data and Knowledge Engineering*, 25:99–123, 1998.

[30] Christine Decaestecker. Finding prototypes for nearest neighbor classification by means of gradient descent and deterministic annealing. *Pattern Recognition*, 30:281–288, 1997.

[31] V. Susheela Devi and M. Narasimha Murty. An incremental prototype set building technique. *Pattern Recognition*, 35:505–513, 2002.

[32] Pierre Devijver. On a new class of bounds on Bayes risk in multihypothesis pattern recognition. *IEEE Transactions on Computers*, 23:70–80, 1974.

[33] Pierre Devijver and Josef Kittler. *Pattern Recognition: A Statistical Approach.* Prectice-Hall, Englewood Cliffs, NJ, 1982.

[34] Pierre A. Devijver and Josef Kittler. On the edited nearest neighbor rule. In *Fifth Int. Conf. Pattern Recognition*, pages 72–80, Miami, Dec. 1980.

[35] Luc Devroye. A universal k-nearest neighbor procedure in discrimination. In *Proceedings of the 1978 IEEE Computer Society Conference on Pattern Recognition and Image Processing*, pages 142–147, 1978.

[36] Luc Devroye. On the inequality of Cover and Hart. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3:75–78, 1981.

[37] Luc Devroye, László Gyorfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag New York, Inc., 1996.

[38] Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. Wiley-Interscience, New York, 1973.

[39] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley and Sons, Inc., New York, 2001.

[40] Sahibsingh A. Dudani. The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, 6:325–327, 1976.

[41] Rex Dwyer. The expected size of the sphere-of-influence graph. *Computational Geometry: Theory and Applications*, 5:155–164, 1995.

[42] Ibrahim Esat. Neural network design based on decomposition of decision space. In *Proceedings of the 6th International Conference on Neural Information Processing*, volume 1, pages 366–370, 1999.

[43] E. Fix and J. Hodges. Discriminatory analysis. Nonparametric discrimination: Consistency properties. Tech. Report 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

[44] K. Florek, J. Lucaszewicz, J. Perkal, H. Steinhaus, and S. Zubrzycki. Sur la liaison et la division des points d'un ensemble fini. *Colloquium Mathematicae*, 2:282–285, 1951.

[45] J. H. Friedman. *Flexible metric nearest neighbor classification*. Stanford University, Stanford, California, November 1994. Technical Report.

[46] Keinosuke Fukunaga and L. D. Hostetler. K-nearest-neighbor Bayes-risk estimation. *IEEE Transactions on Information Theory*, 21:285–293, 1975.

[47] W. Gates. The reduced nearest neighbor rule. *IEEE Transactions on Information Theory*, 18:431–433, 1972.

[48] L. A. Goodman and W. H. Kruskal. Measures of association for cross classifications. *J. of the American Statistical Ass.*, pages 723–763, 1954.

[49] K. Chidananda Gowda and G. Krishna. The condensed nearest neighbor rule using the concept of mutual nearest neighborhood. *IEEE Transactions on Information Theory*, 25:488–490, 1979.

[50] Leo Guibas, Janos Pach, and Micha Sharir. Sphere-of-influence graphs in higher dimensions. In *Intuitive Geometry (Szeged, 1991)*, pages 131–137. North-Holland, Amsterdam, 1994.

[51] Peter E. Hart. The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, 14:515–516, 1968.

[52] Trevor Hastie and Robert Tibshirani. Discriminant adaptive nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18:607–616, June 1996.

[53] K. Hattori and M. Takahashi. A new edited k-nearest neighbor rule in the pattern classification problem. *Pattern Recognition*, 33:521–528, 2000.

[54] Martin E. Hellman and Josef Raviv. Probability of error, equivocation, and the Chernoff bound. *IEEE Trans. Information Theory*, 16:368–372, 1970.

[55] Manabu Ichino and Jack Sklansky. The relative neighborhood graph for mixed feature variables. *Pattern Recognition*, 18:161–167, 1985.

[56] Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Englewood Cliffs, New Jersey, 1988.

[57] Nicholas Jardine and Robin Sibson. *Mathematical Taxonomy*. John Wiley and Sons Ltd, London, 1971.

[58] J. W. Jaromczyk and Godfried T. Toussaint. Relative neighborhood graphs and their relatives. *Proceedings of the IEEE*, 80(9):1502–1517, 1992.

[59] D. Kibler and D. W. Aha. Learning representative exemplars of concepts: An initial case study. In *Proceedings of the Fourth International Workshop on Machine Learning*, pages 24–30, Irvine, CA, 1987.

[60] David G. Kirkpatrick and John D. Radke. A framework for computational morphology. In Godfried T. Toussaint, editor, *Computational Geometry*, pages 217–248. North Holland, Amsterdam, 1985.

[61] T. Kohonen. *Self-Organizing Map*. Springer-Verlag, Germany, 1995.

[62] L. I. Kuncheva and L. C. Jain. Nearest neighbor classifier: Simultaneous editing and feature selection. *Pattern Recognition Letters*, 20:1149–1156, 1999.

[63] Ludmila I. Kuncheva. Fitness functions in editing k-NN reference set by genetic algorithms. *Pattern Recognition*, 30:1041–1049, 1997.

[64] Ludmila I. Kuncheva and J. C. Bezdek. Nearest prototype classification: clustering, genetic algorithms, or random search. *IEEE Transactions on Systems, Man and Cybernetics*, 28:160–164, 1998.

[65] Chang-Hwan Lee and Dong-Guk Shin. Using Hellinger distance in a nearest neighbor classifier for relational data bases. *Knowledge-Based Systems*, 12:363–370, 1999.

[66] U. Lipowezky. Selection of the optimal prototype subset for 1-NN classification. *Pattern Recognition Letters*, 19:907–918, 1998.

[67] Cheng-Lin Liu and Masaki Nakagawa. Evaluation of prototype learning algorithms for nearest-neighbor classifier in application to handwritten character recognition. *Pattern Recognition*, 34:601–615, 2001.

[68] A. Mathai and P. Rathie. *Basic Concepts in Information Theory and Statistics*. Wiley Eastern Ltd., New Delhi, 1975.

[69] K. Matusita. On the notion of affinity of several distributions and some of its applications. *Annals Inst. Statistical Mathematics*, 19:181–192, 1967.

[70] Geoffrey J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley and Sons, Inc., New York, 1992.

[71] F. R. McMorris and C. Wang. Sphere-of-attraction graphs. *Congressus Numerantium*, 142:149–160, 2000.

[72] T. S. Michael and T. Quint. Sphere of influence graphs in general metric spaces. *Mathematical and Computer Modelling*, 29:45–53, 1999.

[73] R. A. Mollineda, F. J. Ferri, and E. Vidal. An efficient prototype merging strategy for the condensed 1-NN rule through class-conditional hierarchical clustering. *Pattern Recognition*, 35:in press, 2002.

[74] O. Murphy, B. Brooks, and T. Kite. Computing nearest neighbor pattern classification perceptrons. *Information Sciences*, 83:133–142, 1995.

[75] Nils J. Nilsson. *The Mathematical Foundations of Learning Machines.* Morgan Kaufmann Publishers, Inc., San Mateo, CA., 1990.

[76] R. Olshen. Comments on a paper by C. J. Stone. *Annals of Statistics*, 5:632–633, 1977.

[77] J. O'Rourke and G. T. Toussaint. Pattern recognition. In Jacob E. Goodman and Joseph O'Rourke, editors, *Handbook of Discrete and Computational Geometry*, chapter 43, pages 797–814. CRC Press LLC, Boca Raton, 1997.

[78] N. R. Pal and J. Biswas. Cluster validity using graph theoretic concepts. *Pattern Recognition*, 30:847–857, 1997.

[79] Y. Park and J. Sklansky. Automated design of multiple-class piecewise linear classifiers. *Journal of Classification*, 6:195–222, 1989.

[80] Y. Park and J. Sklansky. Automated design of linear tree classifiers. *Pattern Recognition*, 23:1393–1412, 1990.

[81] M. S. Paterson and F. F. Yao. On nearest-neighbor graphs. In *Automata, Languages and Programming*, volume 623, pages 416–426. Springer, 1992.

[82] C. S. Penrod and T. J. Wagner. Another look at the edited nearest neighbor rule. *IEEE Transactions on Systems, Man and Cybernetics*, 7:92–94, 1977.

[83] Carey Priebe. *Olfactory classification via randomly weighted nearest neighbors.* Johns Hopkins Univ., Baltimore, Maryland, 1998. Tech. Report 585.

[84] Carey E. Priebe, Jason G. DeVinney, and David J. Marchette. On the distribution of the domination number for random class cover catch digraphs. *Statistics and Probability Letters*, 55:239–246, 2001.

[85] Carey E. Priebe, David J. Marchette, Jason G. DeVinney, and Diego Socolinsky. Classification using class cover catch digraphs. Tech. Report January 15, Johns Hopkins University, Baltimore, 2002.

[86] John D. Radke. On the shape of a set of points. In Godfried T. Toussaint, editor, *Computational Morphology*, pages 105–136. North Holland, 1988.

[87] S. V. Rao and Asish Mukhopadhyay. Fast algorithms for computing β-skeletons and their relatives. *Pattern Recognition*, 34:2163–2172, 2001.

[88] Francesco Ricci and Paolo Avesani. Data compression and local metrics for nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21:380–384, April 1999.

[89] G. L. Ritter, H. B. Woodruff, S. R. Lowry, and T. L. Isenhour. An algorithm for a selective nearest neighbor decision rule. *IEEE Transactions on Information Theory*, 21:665–669, November 1975.

[90] R. Royall. *A Class of Nonparametric Estimators of a Smooth Regression Function*. Stanford University, Stanford, California, 1966. Ph.D. Thesis.

[91] Steven Salzberg, Arthur L. Delcher, David Heath, and Simon Kasif. Best-case results for nearest-neighbor learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:599–608, 1995.

[92] J. S. Sánchez, F. Pla, and F. J. Ferri. On the use of neighborhood-based non-parametric classifiers. *Pattern Recognition Letters*, 18:1179–1186, 1997.

[93] J. S. Sánchez, Filiberto Pla, and F. J. Ferri. Prototype selection for the nearest neighbor rule through proximity graphs. *Pattern Recognition Letters*, 18:507–513, 1997.

[94] J. S. Sánchez, Filiberto Pla, and F. J. Ferri. Improving the k-NCN classification rule through heuristic modifications. *Pattern Recognition Letters*, 19:1165–1170, 1998.

[95] V. Vijaya Saradhi and M. Narasimha Murty. Bootstrapping for efficient handwritten digit recognition. *Pattern Recognition*, 34:1047–1056, 2001.

[96] R. Short and K. Fukunaga. The optimal distance measure for nearest neighbor classification. *IEEE Trans. Information Theory*, 27:622–627, 1981.

[97] Jack Sklansky and Leo Michelotti. Locally trained piecewise linear classifiers. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2:101–111, 1980.

[98] Michael Soss. The size of the open sphere of influence graph in L_∞ metric spaces. In *Proceedings Tenth Canadian Conference on Computational Geometry*, pages 108–109, Montreal, Quebec, Canada, 1998.

[99] Michael Soss. On the size of the Euclidean sphere of influence graph. In *Proceedings Eleventh Canadian Conference on Computational Geometry*, pages 43–46, Vancouver, British Columbia, Canada, 1999.

[100] C. Stanfill and D. L. Waltz. Toward memory-based reasoning. *Communications of the ACM*, 29:1213–1228, December 1986.

[101] C. Stone. Consistent nonparametric regression. *Annals of Statistics*, 8:1348–1360, 1977.

[102] T.-H. Su and R.-C. Chang. On constructing the relative neighborhood graph in Euclidean k-dimensional spaces. *Computing*, 46:121–130, 1991.

[103] Hiroshi Tenmoto, Mineichi Kudo, and Masaru Shimbo. Piecewise linear classifiers with an appropriate number of hyperplanes. *Pattern Recognition*, 31:1627–1634, 1998.

[104] I. Tomek. A generalization of the *k-nn* rule. *IEEE Transactions on Systems, Man and Cybernetics*, 6:121–126, 1976.

[105] I. Tomek. Two modifications of CNN. *IEEE Transactions on Systems, Man and Cybernetics*, 6:769–772, 1976.

[106] Godfried T. Toussaint. Comments on a modified figure of merit for feature selection in pattern recognition. *IEEE Transactions on Information Theory*, 17:618–620, 1971.

[107] Godfried T. Toussaint. Note on optimal selection of independent binary-valued features for pattern recognition. *IEEE Transactions on Information Theory*, 17:618, 1971.

[108] Godfried T. Toussaint. Feature evaluation with quadratic mutual information. *Information Processing Letters*, 1:153–156, 1972.

[109] Godfried T. Toussaint. On information transmission, nonparametric classification, and measuring dependence between randon variables. In *Proc. Symp. Statistics and Related Topics*, pages 30.01–30.08, Ottawa, Oct. 1974.

[110] Godfried T. Toussaint. On some measures of information and their application to pattern recognition. In *Proc. Conf. Measures of Information and Their Applications*, Indian Inst. Technology, Bombay, August 16-18 1974.

[111] Godfried T. Toussaint. On the divergence between two distributions and the probability of misclassification of several decision rules. In *Proc. 2nd International Conf. Pattern Recognition*, pages 27–34, Copenhagen, 1974.

[112] Godfried T. Toussaint. Some properties of Matusita's measure of affinity of several distributions. *Annals Inst. Statistical Mathematics*, 26:389–394, 1974.

[113] Godfried T. Toussaint. A generalization of Shannon's equivocation and the Fano bound. *IEEE Trans. Systems, Man and Cybernetics*, 7:300–302, 1977.

[114] Godfried T. Toussaint. An upper bound on the probability of misclassification in terms of the affinity. *Proceedings of the IEEE*, 65:275–276, 1977.

[115] Godfried T. Toussaint. Probability of error, expected divergence and the affinity of several distributions. *IEEE Transactions on Systems, Man and Cybernetics*, 8:482–485, 1978.

[116] Godfried. T. Toussaint. Algorithms for computing relative neighbourhood graph. *Electronics Letters*, 16(22):860, 1980.

[117] Godfried T. Toussaint. Pattern recognition and geometrical complexity. In *Fifth International Conference on Pattern Recognition*, pages 1324–1347, Miami, December 1980.

[118] Godfried T. Toussaint. The relative neighbourhood graph of a finite planar set. *Pattern Recognition*, 12:261–268, 1980.

[119] Godfried T. Toussaint. A graph-theoretical primal sketch. In Godfried T. Toussaint, editor, *Computational Morphology*, pages 229–260. North-Holland, Amsterdam, Netherlands, 1988.

[120] Godfried T. Toussaint. A counterexample to Tomek's consistency theorem for a condensed nearest neighbor rule. *Pattern Recog. Lett.*, 15:797–801, 1994.

[121] Godfried T. Toussaint. Simple neural networks for nearest-neighbor classification. Tech. Report July, School of Computer Science, McGill University, 2002.

[122] Godfried T. Toussaint, Binay K. Bhattacharya, and Ronald S. Poulsen. The application of Voronoi diagrams to nonparametric decision rules. In *Computer Science and Statistics: The Interface*, pages 97–108, Atlanta, 1985.

[123] Godfried T. Toussaint and Robert Menard. Fast algorithms for computing the planar relative neighborhood graph. In *Proceedings of the Fifth Symposium on Operations Research*, pages 425–428, University of Köln, August 1980.

[124] Godfried T. Toussaint and Ronald S. Poulsen. Some new algorithms and software implementation methods for pattern recognition research. In *Proc. IEEE International Computer Software Applications Conference*, pages 55–63, Chicago, 1979.

[125] R. Urquhart. Graph theoretical clustering based on limited neighborhood sets. *Pattern Recognition*, 15:173–187, 1982.

[126] I. Vajda. A contribution to the informational analysis of pattern. In Satosi Watanabe, editor, *Methodologies of Pattern Recognition*, pages 509–519. Academic Press, New York, 1969.

[127] Jason De Vinney and Carey Priebe. Class cover catch digraphs. *Discrete Applied Mathematics*. in press.

[128] Terry J. Wagner. Convergence of the edited nearest neighbor. *IEEE Transactions on Information Theory*, 19:696–697, September 1973.

[129] Dietrich Wettschereck and David W. Aha. Weighting features. In *Proc. First International Conf. Case-Based Reasoning*, Sesimbra, Portugal, 1995.

[130] Dietrich Wettschereck, David W. Aha, and Takao Mohri. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Review*, 11:273–314, 1997.

[131] Gordon Wilfong. Nearest neighbor problems. In *Proc. 7th Annual ACM Symposium on Computational Geometry*, pages 224–233, 1991.

[132] D. L. Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man and Cybernetics*, 2:408–420, 1972.

[133] D. Randall Wilson and Tony R. Martinez. Instance pruning techniques. In D. Fisher, editor, *Machine Learning: Proc. 14th International Conf.*, pages 404–411. Morgan Kaufmann, San Francisco, 1997.

[134] D. Randall Wilson and Tony R. Martinez. Reduction techniques for instance-based learning algorithms. *Machine Learning*, 38:257–286, 2000.

[135] Charles T. Zahn. Graph theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on Computers*, 20:68–86, 1971.

[136] H. Zhang and G. Sun. Optimal reference subset selection for nearest neighbor classification by tabu search. *Pattern Recognition*, 35:1481–1490, 2002.

[137] Qiangfu Zhao. Stable on-line evolutionary learning of nearest-neighbor multilayer perceptron. *IEEE Trans. Neural Networks*, 8:1371–1378, Nov. 1997.

[138] Qiangfu Zhao and Tatsuo Higuchi. Evolutionary learning of nearest-neighbor multilayer perceptron. *IEEE Trans. Neural Networks*, 7:762–767, May 1996.