Editing Prototypes in the Finite Sample Size Case Using Alternative Neighborhoods

F.J. Ferri[1], J.S. Sánchez[2] and F. Pla[2]

[1] Institut de Robòtica, Universitat de València.
Dr. Moliner 50, 46100 Burjassot, Spain
ferri@uv.es, 34 (9)6 3160 414
[2] Departament d'Informàtica, Universitat Jaume I.
Campus Penyeta Roja, 12071 Castelló. Spain.
{sanchez,pla}@inf.uji.es, 34 (9)64 345676

Abstract. The recently introduced concept of Nearest Centroid Neighborhood is applied to discard outliers and prototypes in class overlapping regions in order to improve the performance of the Nearest Neighbor rule through an editing procedure. This approach is related to graph based editing algorithms which also define alternative neighborhoods in terms of geometric relations. Classical editing algorithms are compared to these alternative editing schemes using several synthetic and real data problems. The empirical results show that the proposed editing algorithm constitutes a good trade-off among performance and computational burden.

1 Introduction

The Nearest Neighbor (NN) rule is one of the most common choices among non-parametric classification rules [2]. In general, given a set of n previously labeled prototypes, the k-NN rule assigns to an unknown sample the label of the majority among its k closest prototypes. Closeness is defined according to a certain measure of dissimilarity, usually the Euclidean distance in a certain feature space. The NN family of rules combine their simplicity in implementation with an appropriate behavior in the expected performance when $n \to \infty$. However, it is well-known that these classifiers suffer from some drawbacks. The first disadvantage comes from the impossibility of having a sufficiently large number of prototypes to achieve (near) optimal performance. On the other hand, large number of prototypes (to approximate the asymptotical performance) implies an important computational burden to find the (k-)nearest neighbor(s) and makes the NN methods inapplicable for problems in which distance calculation is a time-consuming procedure.

A number of different approaches have already been proposed in order to overcome these drawbacks. First, fast searching algorithms [3] or condensed NN rules [4] try to alleviate the computational problem. Second, weighted NN

This work was partially supported by the Spanish CICYT projects TIC95-676-C02-01 and AGF95-0712-C03-01.

rules [9], optimal distance measures [12], and edited NN rules [4] try to improve the performance of the NN rule in different ways. Joint use of edited and condensed NN rules, also referred to as *prototype selection* techniques, lead to a composite approach in which both computational issues and performance are improved [4, 13]. It has been observed by different researchers that asymptotically optimal edited NN rules may lead to arbitrarily bad classification results if the number of prototypes is not large enough compared to the intrinsic dimensionality of the data. This has motivated a number of improvements and alternatives to editing algorithms [5, 7, 10].

The aim of this paper is to present a recent approach to improve the editing rules for small sets of prototypes by using alternative distance measures based on Proximity Graphs (PG) [10] and also, to introduce heuristics about the spatial distribution of the closest prototypes [1] to define new distance measures. These will be referred here to as *surrounding* approaches because they try to look for prototypes close enough (in the basic distance sense) but homogeneously or symmetrically distributed *around* a sample. This idea is inspired in the way the locally optimal distance for NN rules [12] works by using a straightforward approximation of the conditional distribution of prototypes in the neighborhood to account for possible deviations from the asymptotic behavior of the rule. It has also been shown that, as in the case of using optimal distances, the surrounding neighborhoods directly used in the classification rule may constitute an important advantage for finite sample size problems [11].

The paper is organized as follows. Basic concepts about editing algorithms are presented in Section 2. Section 3 includes the definition of alternative neighborhood criteria for its use in editing algorithms. The new editing algorithms making use of these criteria are presented in Section 4. Section 5 includes some of the experiments carried out to compare the different editing schemes considered in this work. Finally, some conclusions and possible extensions are outlined in Section 6.

2 Edited NN rules

The edited NN rule was proposed first by Wilson [14] to improve the performance of the plain NN rule. This editing algorithm can be summarized as follows: any prototype in a given set is discarded if the majority of its k-NN are from a different class. The edited NN rule then consists of applying the 1-NN rule using the edited set of prototypes instead.

It was empirically proven from the beginning that the edited NN rule may outperform the results of the k-NN because it tends to remove outliers and retain only prototypes grouped into clustered regions with no overlapping among classes. However, to strictly prove these facts from an asymptotic point of view required further work and a different (and randomized) way of estimating the true label of a prototype being discarded. All this work lead to the well-known Multiedit algorithm [4]. This tendency to give well clustered and "clean" classes make prototype selection techniques, and editing in particular very appealing from a general point of view because these edited sets may be used to improve or accelerate other learning techniques. In fact, removing outliers is a convenient step prior to applying any classification techniques in most practical cases.

Asymptotically optimal editing algorithms heavily rely on the fact that the estimation of prototype labels have to be done in an statistically independent way. For practical procedures using finite sets, more independence implies less reliability on the estimate. In fact, the Multiedit algorithm performs an internal holdout estimate which is known to exhibit a tendency to overestimate the error as the number of samples decreases [4]. In such cases, the use of a cross-validation estimate [5] or even the leaving-one-out method (Wilson's algorithm) may give better results than optimal editings.

Further improvement may be obtained by looking for neighbors using alternative distance measures or alternative classifier definitions which are known to perform adequately in the finite size case. In this way, the effects from not having infinitely many prototypes is minimized. For example, the distance in [12] has already been used along with the Multiedit-condensing approach [8] but no satisfactory results came out from this combination.

3 Alternative neighborhood definitions

One of the main facts that make the NN rule asymptotically optimal is the convergence of the neighbors to any sample [2]. In this case, any neighborhood around a sample, x, has an empty volume and statistical properties of the sample and all its neighbors are obviously the same. In fact, all asymptotic properties that can be derived for NN rules hold regardless of the distance used.

When there is a finite and small number of prototypes available, the neighborhoods can have arbitrary shape and volume and none of the properties which lead to the asymptotic results holds. In such a situation, Short and Fukunaga [12] proved that an optimal metric in the sense that it locally minimizes the conditional risk difference among the finite and infinite sample cases can be defined. To approximate this distance, the conditional pdf have to be estimated in the neighborhood which results in a serious drawback in practice.

A similar idea consists of considering spatially homogeneous neighborhoods symmetrically distributed around a sample, x. In this way, not all nearest neighbors are considered. If we consider that conditional pdfs are smooth enough in the vicinity of a sample the ratio of prototypes from each class in this *surrounding* neighborhood can give us a better approximation of the conditional aposteriori probabilities which, in practice, improves the k-NN rule in most situations [10]. Surrounding Neighborhoods (SN) can be obtained in a number of ways. First, the special cases of PG as the Delaunay Triangulation (DT), the Gabriel Graph (GG) or the Relative Neighborhood Graph (RNG) can be used to establish a geometrical relation between a sample and some of its neighbors [6, 11].

A new idea has recently come out to obtain a SN based on the concept of Nearest Centroid Neighborhood (NCN) [1] which can be defined as a collection of sufficiently close neighbors whose centroid is also close to the sample x. In practice, the k-NCN can be obtained incrementally as follows:

- the 1-NN equals the 1-NCN.
- the k-NCN can be obtained from the $(k-1)$-NCN by adding the prototype which makes the resulting centroid closest to x.

This iterative procedure clearly does not minimize the distance to the centroid because it gives precedence to the individual distances instead. In this way, the k-NCN makes more sense because it will be very similar to the k-NN but more homogeneously distributed. Even though there is no theoretical framework to account for the NCN concept, the empirical results reported to date are encouraging [1, 11].

4 A Surrounding Editing Technique using the NCN

Most editing algorithms consist of two different but closely related parts: the internal classifier used and the way in which the final label is estimated to decide whether to discard a prototype or not. Although it is possible to use the above neighborhood definitions in any of the classical editing algorithms, only the Wilson editing will be considered in this work to arrive to new editing algorithms. The reasons are two-fold: first because other (holdout based) editings would lead to less computationally feasible solutions and also, because we start from the fact that in the small sample size case the approach from Wilson gives always better results.

The GG and RNG editing would consist of applying the Wilson editing using the graph neighbors of each prototype instead of the k-NN. As a consequence, there is no dependency on k because the number of graph neighbors is determined for each prototype from the same graph. The kind of graphs that could be used are the DT, the GG and the RNG. These three graphs are in fact nested subgraphs which motivates that the number of RNG neighbors is always small compared to the number of GG neighbors. It is not of practical use to consider the DT because all known algorithms to compute it are exponential in the dimension of the feature space, d. Even the GG and RNG require very costly algorithms to compute them [6]. This is a serious drawback for practical applications in which graph computation cannot be done off-line. Also, too sparse graphs (equivalent to low k values) tend to discard a lot of prototypes which may imply a serious degradation in performance.

The NCN concept gives rise to a relatively efficient classifier which can be used with different values of k. The NCN classifier can be used to obtain a more accurate information about prototypes, and specially, for those close to decision boundaries. This will result in a practical improvement of the corresponding editing algorithms. The Wilson algorithm particularized for the case of using the k-NCN would be as follows:

```
k-NCN Editing Algorithm

Step 1  Let $S = X$. (X is the original prototype set)
Step 2  for each $x_i \in X$ do:
    a) Let $T = X - \{x_i\}$.
    b) Find the $k$-NCN of $x_i$ in $T$.
    c) Discard $x_i$ if there is majority of NCNs from a different class.
```

5 Experimental Results

Several experiments using both synthetic and real databases have been carried out in order to compare the efficiency of all editing algorithms considered in this work. Five different random partitions (half for training and half for testing purposes) of each original database. have been used to obtain average measures about the performance of each editing algorithm. In particular. the following approaches have been considered for this experimental study: the Wilson editing [14], PG based editing using both GG and RNG [10] and, finally, the NCN editing. Different settings and versions of the Multiedit and Holdout based editing [4] have been also tried on these databases but the corresponding results are not included in the figures because they were, in general much worse than with the other algorithms. The results corresponding to the k-NN rule have been also included for comparison purposes.

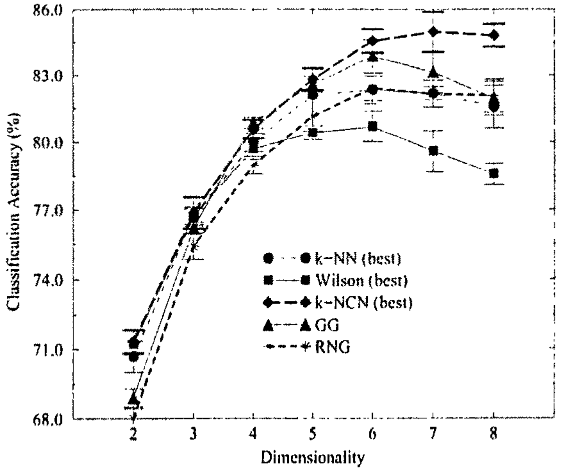

Fig. 1. The synthetic database. Classification accuracy of the different edited NN rules with varying dimensionality.

5.1 The synthetic database

This experiment consists of a set of seven synthetic databases corresponding to the same problem but with dimensionality ranging from 2 to 8. There are two classes consisting of multivariate normal distributions with zero mean and standard deviation 1 and 2 in all dimensions, respectively. There are 2,500 prototypes available in each class. The purpose of this experiment is the study of the behavior of the editing algorithms for different dimensionalities in a problem with strong overlapping among classes. In this experiment only, the results of the k-NN rule, the Wilson editing and the NCN-based editing correspond to the best value of k from all the different neighborhood sizes tried.

As can be seen in Figure 1, all editing procedures gave quite similar performance up to dimension 4. Nevertheless, there is a clear separation in their behavior as the dimensionality increases above 5. All surrounding editings give equal or better results than the k-NN from dimension 6. It is worth mentioning that the Multiedit algorithm gave the best result for this experiment only for the lowest dimensionality (around 73%). The better overall results for higher dimensionalities correspond to the NCN approach. It appears from these results that the NCN editing is less sensitive to the size to intrinsic dimensionality ratio in the training set.

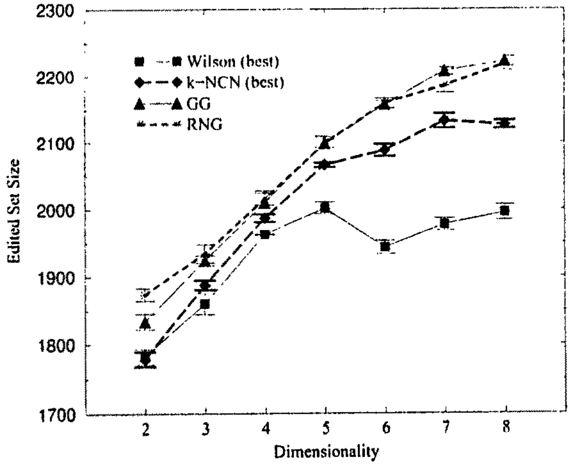

Fig. 2. The synthetic database. Prototypes retained by the different editing algorithms with varying dimensionalities.

Figure 2 shows the averaged number of prototypes retained by each editing procedure along with the corresponding standard deviations. Obviously, all approaches need to retain more prototypes as the dimension increases. Taking into account both Figure 1 and 2, the NCN approach results in a good compromise between classification performance and final edited set size.

5.2 The texture database

This database was generated from the Brodatz's album to study a texture discrimination problem with high order statistics. The aim consists of distinguishing among 11 different textures, each pattern (pixel) being characterized by 40 attributes built from the estimation of fourth order modified moments with four distinct orientations. There is a total of 5.500 prototypes, 500 per class.

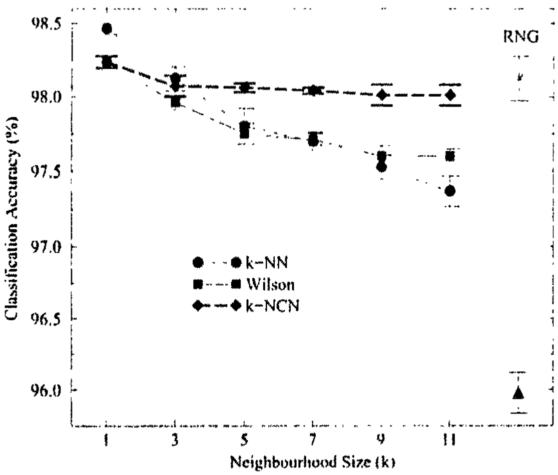

Fig. 3. The texture database. Classification accuracy of the different edited NN rules with various neighborhood sizes.

Figure 3 illustrates the classification rates achieved by the 1-NN rule using the edited sets in this particular problem. Apart from the accuracy of the plain 1-NN rule using the whole training set, all results are quite similar for this problem but the editing using RNG and NCN achieved the best accuracy levels.

With regard to the size of the edited sets, shown in Figure 4, it can be seen that all methods retain a very similar number of prototypes apart from the GG-based editing which retain a fraction of prototypes similar to the one obtained by the Multiedit algorithm for this problem (about 80%).

5.3 The Landsat image database

The purpose of the third experiment is the classification of the multi-spectral values of a real image (2,340 x 3,380 pixels) from the Landsat satellite. This database results from a sub-area of an image, consisting of 82 x 100 pixels.

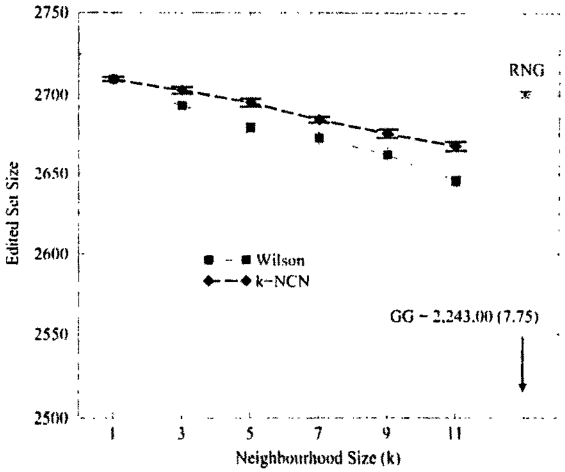

Fig. 4. The texture database. Averaged number of prototypes retained by different editing algorithms.

Each pattern corresponds to a 3x3 square neighborhood of pixels completely contained within that sub-area. There are 6,435 samples with 36 attributes (4 spectral bands x 9 pixels in the neighborhood) and six classes.

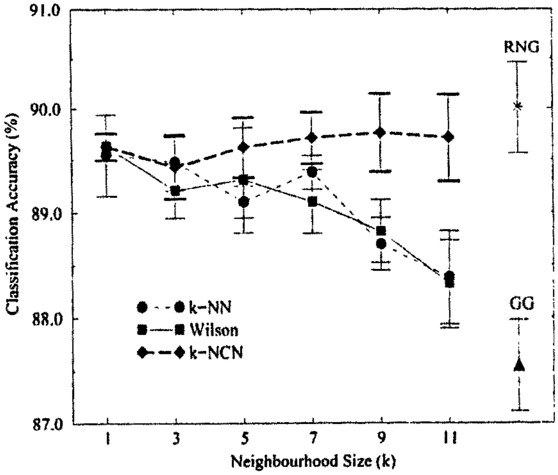

Fig. 5. The landsat database. Classification accuracy of the different edited NN rules with various neighborhood sizes.

From Figure 5, it seems clear that the surrounding approaches achieve the highest classification rates and, in particular, the NCN based approach is more stable as the neighbor size increases. The RNG based editing gives the best accuracy level for this problem but the differences between this and the NCN based editing are not statistically significant.

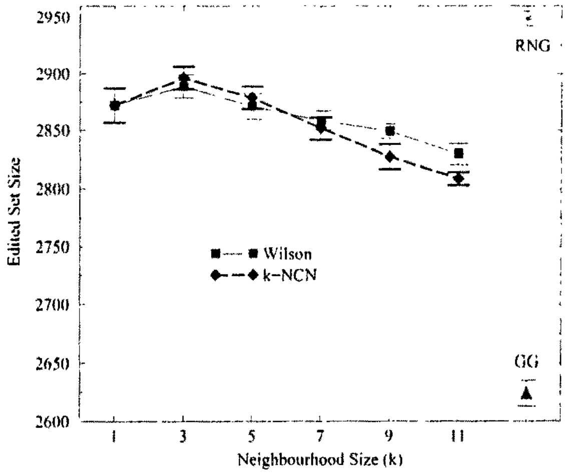

Fig. 6. The landsat database. Averaged number of prototypes retained by different editing algorithms.

Figure 6 shows the averaged number of prototypes retained for this problem. From this, it is quite clear that the RNG requires significantly more prototypes than the other algorithms. Also, both Wilson and NCN based editing exhibit a decreasing tendency on k which gives rise again to a good trade-off between accuracy and final set size for the NCN based editing.

6 Concluding Remarks and Further Work

Alternative approaches to editing the NN rule have been considered in this work. In particular, the recently introduced concept of NCN applied to editing has been proposed. In general, the so-called surrounding approaches to editing look for close and symmetrically distributed prototypes to decide about retaining a prototype or not. In this sense, the NCN provides an efficient and convenient way of obtaining such a neighbors. From the experiments carried out it can be concluded that either the RNG or the NCN based editing are the best options to obtain appropriately edited sets. Moreover, the NCN retain significantly less prototypes and, more importantly, it does not require as much computation as the PG based approach. In fact, brute force computation of the k-NCN requires

$O(kn)$ while using an heuristic optimization to compute the PG results approximately in $O(dn^2)$, where d is the dimensionality of the feature space.

The results obtained in the experiments carried out in this work are encouraging enough to further continue studying surrounding approaches to editing. New alternative neighborhoods, possibly better connected to the asymptotic case under a convenient theoretical framework, could be possible. Although it is possible to use SN in other classical editing schemes as the Holdout editing and Multiedit, it is not worth it mainly because that would require recomputing graphs or the NCN for each block and for each iteration of the algorithm, respectively. A better way of proceeding would be to combine information about different neighborhoods into a single editing algorithm.

References

1. B.B. Chaudhuri. A new definition of neighbourhood of a point in multi-dimensional space. *Pattern Recognition Letters*, 17:11–17, 1996.
2. T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.
3. B. V. Dasarathy and B. V. Sheela. Visiting nearest neighbors. In *Proc. Int. Conf. on Cybernetics and society*, pages 630–635, 1977.
4. P. A. Devijver and J. Kittler. *Pattern Recognition. A Statistical Approach.* Prentice Hall, 1982.
5. F. Ferri and E. Vidal. Small sample size effects in the use of editing techniques. In *Proc. of 11th International Conference of Pattern Recognition*, pages 607–610, The Hague, THE NETHERLANDS, September 1992.
6. B.K. Bhattacharya G.T. Toussaint and R.S. Poulsen. The application of voronoi diagrams to nonparametric decision rules. In L. Billard, editor, *Computer Science and Statistics: The Interface*. Elsevier Science, North-Holland, 1985.
7. L. Kuncheva. Editing for the k-nearest neighbors rule by a genetic algorithm. *Pattern Recognition Letters*, 16(8):809–814, 1995.
8. A.E. Lucas and J. Kittler. A comparative study of the kohonen and multiedit neural net learning algorithms. In *Proc. 1st IEE Int. Conf.on Artificial Neural Networks*, pages 7–11, 1991.
9. J.E.S. Macleod, A. Luck, and D.M. Titterington. A re-examination of the distance-weighted k-nearest-neighbor classification rule. *IEEE Transactions on Systems Man and Cybernetics*, 17(4):689–696, 1987.
10. J.S. Sánchez, F. Pla, and F.J. Ferri. Prototype selection for the nearest neighbour rule through proximity graphs. *Pattern Recognition Letters*, 18(7):507–513, 1997.
11. J.S. Sánchez, F. Pla, and F.J. Ferri. On the use of neighbourhood-based non-parametric classifiers. *Pattern Recognition Letters*, (in press), 1998.
12. R.D. Short and K. Fukunaga. The optimal distance measure for nearest neighbor classification. *IEEE Transactions on Information Theory*, 27(5):622–627, 1981.
13. J. Voisin and P. A. Devijver. An application of the multiedit-condensing technique to the reference selection problem in a print recognition system. *Pattern Recognition*, 20(5):465–474, 1987.
14. D. L. Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems Man and Cybernetics*, 2(3):408–421, 1972.