ELSEVIER

Pattern Recognition Letters 16 (1995) 809–814

Pattern Recognition Letters

Editing for the k-nearest neighbors rule by a genetic algorithm

Ludmila I. Kuncheva *

Department of Biomedical Engineering (CLBME), Bulgarian Academy of Sciences, Acad. G. Bonchev Street, Block 105, 1113 Sofia, Bulgaria

Received 28 September 1994; revised 1 March 1995

Abstract

A genetic algorithm is applied for selecting a reference set for the k-Nearest Neighbors rule. The performance has been evaluated on a medical data set by the rotation method. The results are commented together with those obtained with the standard k-NN, random selection, Wilson's technique, and the MULTIEDIT algorithm.

Keywords: k-NN rule; Editing techniques; Genetic algorithms

1. Introduction

Due to its robustness, theoretical elegance, and feasibility of realization, the k-Nearest Neighbors (k-NN) rule continues to be one of the most widely used classification techniques. The efforts to select an optimal reference subset (called also a design set of prototypes) from an initial data set have stemmed from the need to reduce the immense storage requirements and computational loads connected with the rule. The second perspective on this subject, as pointed out by Dasarathy (1990), is, by editing out some objects from the sample, to achieve a classification result that is hopefully more accurate and reliable. The two perspectives are not clearly distinguishable, and in fact are largely merged in many studies. The emphasis in this paper is on the classification accuracy.

Numerous editing techniques are summarized by Dasarathy (1990, 1994) and Devijver and Kittler (1982). Three general classes can be formulated (a technique may be attributed to more than one class):

Condensed Nearest Neighbor rule (CNN). The numerous studies on CNN commenced with Hart's paper (1968) that surprisingly ends up with a negative experimental conclusion. Nevertheless it has given rise to many other editing strategies and techniques. Under this name we will assume all editing techniques the result of which is a subset of the original reference set: neither are new prototypes generated, nor are any data items modified.

Generated or modified prototypes. This group comprises techniques that either establish new prototypes (see Marin and Mira, 1991) or adjust a limited number of points from the initial data set. A large group of such techniques are implemented by neural networks, e.g. feature-map classifiers, learning vector quantiziers, hypersphere classifiers, etc. (see Lippmann, 1989; Moed and Lee, 1993; Reilly et al.,

* Email: lucy@bgcict.bitnet

SSDI 0167-8655(95)00047-X

1987; Yau and Manry, 1991; Decaestecker, 1993; Holmstrom and Hamalainen, 1993; Katz and Thrift, 1993).

Two-level classifiers (Combination of multiple classifiers). This group contains several heuristic solutions to the editing problem employing a two-level classification strategy. One of these is to use the k-NN rule only in a limited subspace of the initial feature space where this technique appears most appropriate, while another classification rule is responsible for the rest of the feature space (Dasarathy, 1990). This automatically leads to a smaller number of reference points to be stored and hopefully yields a more accurate classification decision (Kuncheva, 1993). Alpaydin (1993) proposes a two-level scheme consisting of several neural networks (decision makers), each one performing the 1-NN rule on a reference set selected by the original Hart's procedure with different arrangements of the initial set.

In this paper a genetic algorithm (GA) has been chosen as an editing technique. The scheme belongs to the first group, since it is designed to select a subset of the original set of reference objects. Unlike in the work of Kelly and Davis (1991) where a GA has been applied to adjust the distance function for a k-NN rule, here the chromosome is directly mapped onto the reference set. The rationale for choosing a GA as an editing technique is the following. It has been proven (see Devijver and Kittler, 1982) that an optimal editing strategy would ultimately retain the objects belonging to the Bayesian decision regions of their own class. In a finite-sample case it is not a priori clear which of the objects satisfy the above condition. We base our choice on the intuition that a reference subset of objects which provides the highest classification accuracy (assessed on the available set) will outperform the initial set in its capacity of the reference set.

Experiments with real medical data have been carried out in order to investigate the classification accuracy of the edited reference set. The rotation (π-method) has been used to assess the classification performance. The results with the GA compare favorably to those with the unedited reference set and with random selection, although the differences are not statistically significant. A comparison with Wilson's (1972) editing technique and with the MULTIEDIT algorithm (Devijver and Kittler, 1980, 1982) is also reported.

2. Competing editing strategies

First of all we require that our edited reference set yields a better result than using the whole reference set. Second, the GA selection must be better than the random selection of subsets.

Furthermore, as a benchmark, we chose Wilson's (1972) editing technique since it is easily implementable, and does not depend on the order of the objects in the initial set. The technique consists in: first designating for deletion the objects that have been misclassified by a k-NN rule and then removing them from the sample. The rest is used as a reference set and the 1-NN rule is applied for further classification. Indeed, Wilson's estimate has been shown to be biased. This may lead to overestimation of the classification accuracy on the training sample and, consequently, to a lower generalization ability. In order to avoid this, the MULTIEDIT algorithm has been proposed by Devijver and Kittler (1980, 1982). The repeated editing strategy has been proven to be asymptotically Bayes-optimal with respect to the original classification problem. The algorithm follows directly the idea of editing out the objects being misclassified. In contrast to Wilson's technique, this one uses an *independent* reference sample to attach a class label to an object and thus to decide its fate. The algorithm consists in the following steps.

Step 1: *Diffusion.* Let $\mathbf{Z} = \{Z_1, \ldots, Z_n\}$ be the set of reference objects with known class labels. Make a random partition on $\mathbf{Z}$ into q subsets, $\mathbf{Z}^1, \ldots, \mathbf{Z}^q$ ($q \geqslant 3$).

Step 2: *Classification.* Classify the samples in $\mathbf{Z}^i$ using the 1-NN rule with $\mathbf{Z}^{(i+1) \bmod q}$ as a reference set.

Step 3: *Editing.* Discard all the samples that were misclassified at Step 2.

Step 4: *Confusion.* Pool all the remaining data to constitute a new set $\mathbf{Z}$.

Step 5: *Termination.* If the last I iterations produced no editing, exit with the final set $\mathbf{Z}$, else go to Step 1.

3. GA as an editing procedure

Genetic algorithms are a powerful tool to search in a high-dimensional space. Imitating the process of natural reproduction they are a kind of guided random search. The procedure operates on several candidate points simultaneously. One of their most appealing advantages is that they do not require the derivative of the search criterion, nor its continuity.

Let $\Im(\boldsymbol{Z})$ be the power set of $\boldsymbol{Z}$. The problem of selecting the best reference set is formally stated here as: Find $Z^* \in \Im(\boldsymbol{Z})$ such that

$$J(Z^*) = \max\{J(Y) \mid Y \in \Im(\boldsymbol{Z})\}.$$

In order to do the search with a GA, we map the set Y, a subset of $\boldsymbol{Z}$, onto a chromosome structure in the following way. The chromosome consists of n genes, each one with two possible states: 0 and 1. A set Y is presented straightforwardly by assigning the ith gene the value 1 if Z_i is included in Y and 0, otherwise.

The first criterion (fitness function) tried in this investigation was the apparent error rate measured by a *pseudo*-leave-one-out application of the k-NN rule. Like in Wilson's technique, the leave-one-out procedure is not implemented in its pure form because the objects, whose classification is to be assessed, may have participated in the formulation of the reference set.

In more detail, let $Y \subseteq \boldsymbol{Z}$ be the current subset-candidate. The value of $\boldsymbol{J}(Y)$ is measured as $\boldsymbol{n}(Y)/n$ where $\boldsymbol{n}(Y)$ denotes the number of correctly classified objects from $\boldsymbol{Z}$ using only Y to find the k nearest neighbors. For each object Z_i in Y the k nearest neighbors are searched among those in the set $Y \setminus \{Z_i\}$.

Let $Z^c(Y) \subseteq \boldsymbol{Z}$ denote the set of the correctly classified objects from $\boldsymbol{Z}$ under the current Y, and let k_j, $j = 1, \ldots, \boldsymbol{n}(Y)$ be the number of neighbors leading to the correct classification of the object $Z_j \in Z^c(Y)$. We choose the second criterion function which underlies a second session of experiments $\boldsymbol{J}(Y)$ to be

$$\boldsymbol{J}(Y) = \frac{1}{n} \sum_{j=1}^{\boldsymbol{n}(Y)} k_j.$$

The correctly classified patterns with low classification score for their own class will contribute in a lower degree to the value of the criterion. According to this criterion, the GA is expected to converge to a chromosome corresponding to a reference set which assures the highest possible "certainty" of the classification decision. The algorithm does not necessarily retain only points lying deep inside the Bayes decision regions, nor is it purposefully instructed to select boundary points. The decision about the balance is left to the searching procedure. If the first criterion can be viewed as a *pseudo*-leave-one-out counting estimator, the second one is a kind of smoothing modification.

In brief, the genetic algorithm used here consists of the following steps.

1. Generate an initial population set $\Pi = \{Y_1, \ldots, Y_{ps}\}$ consisting of ps chromosomes, where ps is a preliminary determined population size (for the current experiments $ps = 50$). Calculate the fitness values of the chromosomes.
2. Set the current number of generations $i = 1$.
3. Form a mating set $\boldsymbol{M}$ from all the chromosomes with fitness function above the average.
4. Select randomly couples of parents. Choose a random crossover point and exchange the right parts of the parents' chromosomes, thus producing two offspring chromosomes. Put the offsprings in the set $\boldsymbol{O}$. The crossover probability was 1.0 in the current experiments, which means that every selected couple of parents will result in two new chromosomes.
5. Mutate each gene of each offspring in $\boldsymbol{O}$ with a preliminary defined probability (in our setting the mutation rate was 0.05). Calculate the offsprings fitness function.
6. Combine Π and $\boldsymbol{O}$ selecting the best ps chromosomes from their union (elitist strategy). Consider the result as a new population set Π.
7. If i is less than the preliminary selected terminal number of generations then increase i and GO TO 3.

As a result we obtain a population set Π^* that contains ps subsets of $\boldsymbol{Z}$ with the highest values of the criterion function. Since we have not put any restriction trying to keep a diversity in the population, the final chromosomes might appear "relatives", i.e., highly overlapping. Theoretically they can even be ps copies of one unique solution. It is

expected that the GA will find a sufficiently good solution, i.e., it could select reference subsets with sufficiently high classification accuracy. The point of interest is how robust this selection is with respect to generalization. In order to investigate this, the rotation testing method has been used as described in Section 4.

It is worth mentioning that the proposed editing technique is not limited to two-class problems, nor is it confined to a certain number of k. Although it has been proven that 1-NN repeated editing with ensuing 1-NN classification is asymptotically Bayes-optimal (Devijver and Kittler, 1982), there is no guarantee that this optimality is valid for the finite-sample case. Therefore, we used in the experiments the 1-NN, 3-NN, and 9-NN techniques.

4. Statement of the experiments

A data set from aviation medicine has been used to test the proposed editing technique. The task is to predict if a pilot will exhibit cardiac rhythm disorders during a centrifuge training. The examination is carried out in a centrifuge cabin under a profile of high $+Gz$ radial accelerations simulating aerial combat maneuvering. There is some evidence that this kind of training may provoke serious ectopy in asymptomatic healthy people. Such a disorder may also occur in a real flight and cause a fatal accident. An issue that has been investigated for more than a year is the possibility to predict the occurrence of this event on the basis of some anthropometric and physiologic parameters of the pilot measured immediately before the examination. The classification accuracies obtained so far are quite discouraging (see Kuncheva and Zlatev, 1994). It appeared that the two classes (pilots with and without extrasystoles) are hardly distinguishable and almost all classification techniques yield results quite close to the a priori probabilities. The reason to choose this particular data set was that highly overlapping classes may be better material to study the merits of editing procedures than data sets with well separable classes.

Five parameters are used:

- Age of the pilot,
- Height of the pilot,
- Systolic blood pressure immediately before the examination,
- Diastolic blood pressure immediately before the examination,
- Heart rate immediately before the examination.

The two aforementioned classes are considered. The available set contains 485 preclassified records. This set was randomly divided into five nonoverlapping subsets, each one containing 97 objects, used as the training sets and the rest 388 cases used as the independent test sets. The results have been averaged over the five samples. The same procedure was also performed with the MULTIEDIT algorithm in order to avoid an eventual bias of the estimate caused by the random partition of the set at each iteration. We selected $q = 3$ because of the small training-sample size.

5. Results and discussion

As in some previous studies, the results show that the two classes of pilots are not easily distinguishable in the considered feature space. Since we do not favor the decision for the second class, it may turn out that the best overall classification accuracy is achieved by classifying all the objects in the first

Table 1
Averaged results with the five test sets [% correct]

Method for reference set design	Classification accuracy	95% confidence intervals
Whole sample (1-NN test)	63.00	58.20, 67.80
Whole sample (3-NN test)	67.74	63.09, 73.39
Whole sample (9-NN test)	72.30	67.85, 76.75
Random selection (3-NN)	66.21	61.50, 70.92
Random selection (9-NN)	72.59	68.15, 77.03
Wilson's technique (3-NN training, 1-NN test)	72.26	67.80, 76.71
MULTIEDIT (multiple 1-NN training, 1-NN test)	74.44	70.10, 78.78

(more probable) class. The estimated a priori probabilities for class "*pilots without rhythm disorders*" and "*pilots with rhythm disorders*" are 0.7444 and 0.2556, respectively.

The results from the k-NN rule with the whole sample (unedited) are shown in Table 1 with $k = 1$, 3, and 9. Averaged results with randomly selected reference sets (10 per each division into training and test sets) along with the results from Wilson's and MULTIEDIT techniques are also reported in Table 1. The 95% confidence intervals of the provided estimates are also presented.

The results from the GA selection of a reference set with 9-NN are presented in Fig. 1. Only the test sample is considered for estimating the classification accuracy, as in the above experiments. The averaged classification accuracy is depicted versus the generation number. The top picture shows the accuracy averaged over the whole population, and the one below, the accuracy with the best chromosome, as ordered by the fitness function (on the training set).

Contrary to the expectations, the test performance with the first criterion deteriorates with increasing generation number. This means that the first criterion, although the most natural one, bears a high risk of leading to high memorization versus low generalization ability of the classifier. The second criterion yielded better results. Since the other (competing) strategies do not depend on the number of generation, they are presented as straight lines in the figure. After 3–7 iterations, the MULTIEDIT algorithm ruled out all objects from the second class, thus leading to the classification with the a priori probability. This means that the algorithm did not find any "compact" clusters from the second class, worth to be retained. The final sets obtained with the GA, although with nearly the same classification accuracy, contained objects from both classes.

It is clear that the GA converged to coherent final solutions because the maximal and the average population rates are practically the same. Therefore, any of the chromosomes in the final population may serve as the reference set (if there are different individuals at all).

The results with the GA editing technique appeared quite comparable with those from Wilson's

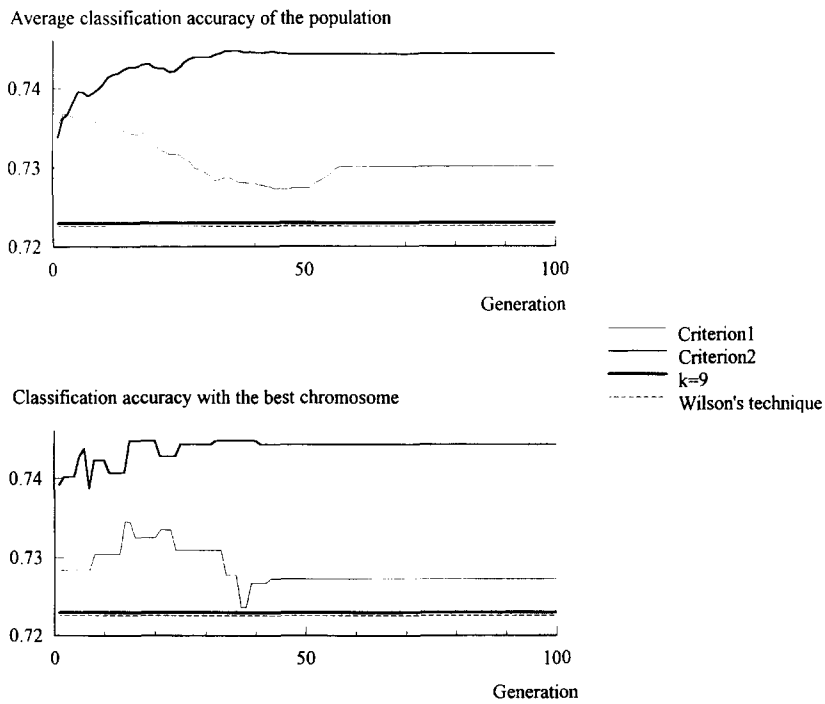

Fig. 1.

technique, and better than those of the classical settings. It was not possible to properly evaluate the power of the MULTIEDIT algorithm with the current data set, and therefore it cannot be contrasted with the proposed technique. It should be kept in mind that, since the differences are not statistically significant, the above comparison is only illustrative and is not meant to state a definitive priority of one technique over the others.

6. Conclusions

A genetic algorithm is proposed as an editing technique for the k-NN rule. Two criteria have been used as the fitness function: the apparent error rate, and a criterion based on the certainty of the classification. In result, it appeared that the second criterion selected a population with chromosomes corresponding to subsets of the initial set that provide higher classification accuracy in comparison with the whole initial set, with random selection and with Wilson's technique.

Acknowledgments

The author wishes to express her sincere thanks to Dr. Roumen Zlatev from the Institute for Aviation Medicine, Military Medical Academy, Sofia, for providing the data set. The kind help of Mr. Yordan Yotzov from the New Bulgarian University, who designed the software and conducted the GA experiments, is highly appreciated.

References

Alpaydin, E. (1993). Multiple networks for function learning. *Proc. IEEE Internat. Conf. on Neural Networks*, San Diego, CA, 9–14.

Dasarathy, B.V. (1990). *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. IEEE Computer Society Press, Los Alamitos, CA.

Dasarathy, B.V. (1994). Minimal consistent set (MCS) identification for optimal nearest neighbor decision systems design. *IEEE Trans. Syst. Man Cybernet.* 24, 511–517.

Decaestecker, C. (1993). NNP: A neural net classifier using prototype. *Proc. IEEE Internat. Conf. on Neural Networks*, San Francisco, CA, 822–824.

Devijver, P.A. and J. Kittler (1982). *Pattern Recognition: A Statistical Approach*. Prentice-Hall Internat., London.

Devijver, P.A. and J. Kittler (1980). On the edited nearest neighbor rule, *Proc. 5th Internat. Conf. on Pattern Recognition*, 72–80.

Hart, P.E. (1968). The condensed nearest neighbor rule. *IEEE Trans. Inform. Theory* 16, 515–516.

Holmstrom, L. and A. Hamalainen (1993). The self-organizing reduced kernel density estimator. *Proc. IEEE Internat. Conf. on Neural Networks*, San Francisco, CA, 417–420.

Katz, A. and P. Thrift (1993). Hybrid neural network classifiers for automatic target detection. *Expert Systems* 10, 243–250.

Kelly, J.D. and L. Davis, Jr. (1991). Hybridizing the genetic algorithm and the k-nearest neighbors classification algorithm. *Proc. 4th Internat. Conf. on Genetic Algorithms*, San Diego, CA, 377–382.

Kuncheva, L.I. (1993). 'Change-glasses' approach in pattern recognition. *Pattern Recognition Lett.* 14, 619–623.

Kuncheva, L.I. and R.Z. Zlatev (1994). Prediction of cardiac disorders of pilots during centrifuge training using a model of fuzzy neuron. In: R. Trappl, Ed., *Cybernetics and Systems II*. World Scientific, Singapore, 991–998.

Lippmann, R.P. (1989). Pattern classification using neural networks. *IEEE Communication Magazine*, 47–64.

Marin, R. and J. Mira (1991). On knowledge-based fuzzy classifiers: A medical case study. *Fuzzy Sets and Systems* 44, 421–430.

Moed, M. and C.-P. Lee (1993). Design of an elliptical neural network with application to degraded character classification. *Proc. Internat. Conf. on Neural Networks*, San Diego, CA, 1576–1582.

Reilly, D.L., C. Scofield, C. Elbaum and L. Cooper (1987). Learning system architectures composed of multiple learning modules. *Proc. 1st Internat. Conf. on Neural Networks*, San Diego, CA, II-495–503.

Wilson, D.L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Trans. Syst. Man Cybernet.* 2, 408–421.

Yau, H.C. and M.T. Manry (1991). Iterative improvement of a nearest neighbor classifier. *Neural Networks* 4, 517–524.