ELSEVIER

Pattern Recognition Letters 19 (1998) 1165–1170

Pattern Recognition Letters

Improving the *k*-NCN classification rule through heuristic modifications [1]

J.S. Sánchez [a,*], F. Pla [a], F.J. Ferri [b]

[a] *Dept. d'Informàtica, Universitat Jaume I, E-12071 Castelló, Spain*
[b] *Dept. d'Informàtica i Electrónica, Universitat de València, E-46100 Burjassot (València), Spain*

Received 14 April 1998; received in revised form 6 August 1998

Abstract

This paper presents an empirical investigation of the recently proposed *k-Nearest Centroid Neighbours* (*k*-NCN) classification rule along with two heuristic modifications of it. These alternatives make use of both proximity and geometrical distribution of the prototypes in the training set in order to estimate the class label of a given sample. The experimental results show that both alternatives give significantly better classification rates than the *k*-Nearest Neighbours rule, basically due to the properties of the plain *k*-NCN technique. © 1998 Published by Elsevier Science B.V. All rights reserved.

Keywords: Nearest centroid neighbourhood; Nearest neighbourhood; Classifier

1. Introduction

The *k-Nearest Neighbours* (*k*-NN) rule (Duda and Hart, 1973) is one of the most remarkable choices among non-parametric classification rules. This is a distance-based technique which classifies a test sample according to the classes of its k closest cases in a set of n previously labelled prototypes, $X = \{x_1, \ldots, x_n\}$. It is well known that the error for the *k*-NN rule tends towards the *Bayes* error in the asymptotic case ($n \to \infty$). However, in practice, due to the finite sample size, the *k*-NN estimates are no longer optimal. This problem becomes more relevant when the number of prototypes is not large enough compared to the dimensionality of the feature space (Fukunaga, 1990), which constitutes a very usual practical situation.

A number of alternative neighbourhood definitions have been applied to classification problems, trying to partially overcome the practical drawbacks pointed out for the *k*-NN rule. In particular, the concept of *Nearest Centroid Neighbourhood* (NCN) (Chaudhuri, 1996) along with the neighbourhood relation derived from the *Gabriel* and the *Relative Neighbourhood* graphs (Jaromczyk and Toussaint, 1992) have successfully been used in finite sample size situations (Sánchez et al., 1997). The resulting classification approaches have been generically referred to as *surrounding* rules because they try to look for prototypes not only close enough (in the basic distance sense) but also homogeneously or symmetrically distributed *around* a sample.

* Corresponding author. E-mail: sanchez@uji.es.

[1] This work has partially been supported by projects P1B96-13 (Fundació Caixa-Castelló), and AGF95-0712-C03-01 and TIC95-676-C02-01 (Spanish CICYT).

0167-8655/98/$ – see front matter
PII: S0167-8655(98)00108-1

Although the surrounding classification schemes have been proven to outperform the k-NN rule in most cases, this kind of neighbourhood also suffers from some drawbacks due to the fact that it may contain some prototypes which are not sufficiently close to the test sample. Thus, this paper proposes some modifications of the k-NCN rule which try to solve this problem and then to achieve better results.

The organization of the rest of this paper is as follows. Section 2 describes the NCN concept and the derived k-NCN classifier, as well as the conceptual differences with respect to the k-NN rule. In Section 3, two modifications of the k-NCN rule are introduced. Section 4 provides an experimental study for both synthetic and real data sets. Finally, some concluding remarks are given in Section 5.

2. Surrounding neighbourhood

The k-NN rule consists of estimating the class of a given sample through its k closest prototypes in the training set. This technique considers that all the information required to classify a new sample can be obtained from a small subset of prototypes close to it. However, it does not take into account the geometrical distribution of those k prototypes with respect to the given sample, that is, in general the nearest prototypes do not completely surround the sample since the k-NN rule defines the neighbourhood only in terms of minimum (Euclidean) distance.

In the asymptotic case, the convergence of the neighbours to a given sample holds regardless of the distance used. Nevertheless, when a finite and small number of prototypes is available, the neighbourhoods can have arbitrary shape and volume and none of the asymptotic properties holds. In such a situation, it has been proven that the use of local distance measures can significantly improve the behaviour of a classifier (Short and Fukunaga, 1980).

The non-parametric k-NCN decision rules along with those based on the proximity graphs aforementioned (Sánchez et al., 1997) also estimate the class of an unknown sample from its neighbours, but using an alternative definition of neighbourhood (the so-called *surrounding neighbourhood*) which allows to focus on a sufficiently small region of prototypes homogeneously distributed *around* the given sample. From all distinct surrounding realizations, the NCN approach provides an efficient and convenient way of obtaining such a neighbourhood. In fact, this concept gives rise to the most suitable surrounding classification scheme (Sánchez et al., 1997), not only because of its excellent performance but also because of its moderate computational cost, which results in O(kn) (Chaudhuri, 1996).

2.1. The nearest centroid neighbourhood

The NCN concept (Chaudhuri, 1996) is based on the idea that the neighbourhood of a point is subject to two constraints. First, by the *distance criterion*, the k neighbours of a sample, p, must be as near as possible. Second, by the *symmetry criterion*, their centroid must be also as close to p as possible. Note that the conventional *nearest neighbourhood* takes into account the first property only, and nearest neighbours may not be symmetrically distributed around p.

Let p be a point whose k nearest centroid neighbours should be found in a set of points (training set), $X = \{x_1, \ldots, x_n\}$. These k neighbours can be searched for through an iterative procedure (Chaudhuri, 1996) in the following way:

1. The first neighbour of p is also its nearest neighbour, q_1.
2. The ith neighbour, q_i, $i \geqslant 2$, is such that the centroid of this and previously selected neighbours, $q_1, \ldots, q_i$, is the closest to p.

This procedure gives rise to a kind of neighbourhood in which the spatial distribution of neighbours is taken into account because of the centroid criterion. On the other hand, proximity of the k nearest centroid neighbours to the sample is guaranteed due to the incremental nature of the way in which they are obtained from the first nearest neighbour.

2.2. The k-NCN classification rule

The aim of estimating the class of a sample from its surrounding neighbours can be accom-

plished by using the NCN concept. In concrete, this classifier, called k-NCN (Sánchez et al., 1997), tries to acquire a generally more reliable result than that obtained by means of the conventional NN definition. Thus, it is proposed to make a decision about the class membership of a given sample after knowing the spatial distribution of the prototypes around it. This means that a sample is classified not only according to its nearest neighbours but also taking into account how prototypes are placed around it.

Given a set of prototypes $X = \{x_1, \ldots, x_n\}$ from L different classes, and a new sample q, the k-NCN classification scheme is defined as follows:

1. Find the k nearest centroid neighbours of q.
2. Assign to q the class with a majority of votes among its k nearest centroid neighbours (resolve ties randomly).

3. Using proximity and spatial homogeneity for classification

We here propose two heuristic modifications of the k-NCN decision rule in order to improve its correct classification rate. These alternative schemes try to jointly use information about proximity as well as about the spatial distribution of prototypes around a given sample. In fact, although it has been empirically proven that the k-NCN rule may outperform the k-NN classifier (Sánchez et al., 1997), some nearest centroid neighbours may be too far from the sample to classify, which can constitute a relative practical drawback. Thus, the aim of these approaches will be to overcome this problem by means of the classical NN concept, trying to get a good trade-off between the nearest neighbourhood and the surrounding neighbourhood.

3.1. Averaged k-NCN rule

The first extension to the k-NCN rule is basically meant to resolve the k-NCN ties. Thus, the tie-breaking modification can be defined as follows: a test sample is assigned to the class which has received at least half of the k votes (say, $l = k/2$). If none of the possible classes receives at least l votes, the average distance between the test sample and the nearest centroid neighbours from a same class is computed, in a way similar to the strategy proposed in (Rabiner et al., 1979). In this latter case, the sample will be finally assigned to the class with the least average distance among all classes voted.

This alternative is similar to that defined for the (k,l)-NN rule, that is, the k-NN rule with a reject option (Devijver and Kittler, 1982). However, it mainly differs from the (k,l)-NN rule in that the reject option is here replaced by an additional distance test. The classification scheme would be as follows:

1. Find the k nearest centroid neighbours of q.
2. If there exists a class, ω_i $(i = 1, 2, \ldots, L)$, with at least $l = k/2$ votes, then assign q to such a class.
3. If not, compute the average distance between q and each one of the classes voted. Assign q to that class with the least average distance.

The algorithm combines the surrounding neighbourhood with the proximity information in "doubtful" cases only. Thus, it is also possible to use other settings for the parameter l different from $k/2$. This would simply mean to impose a more or less stringent majority on the number of votes. On the other hand, note that the modified rule with $l = k/2$ is equivalent to the plain k-NCN rule when $L = 2$.

3.2. Edited k-NCN rule

The second proposal corresponds to editing the set of prototypes (Dasarathy, 1990) since it just considers those nearest centroid neighbours which have not been mislabelled. Thus, after computing the k nearest centroid neighbours of a test sample, we only select those whose nearest neighbour among all prototypes in the training set has their same class label assigned. The advantage of using an editing strategy comes from the fact that it tends to remove outliers and retain only prototypes grouped into clustered regions with no overlapping among classes.

We obtain a classification scheme which combines the surrounding neighbourhood provided by the k-NCN rule with the conventional nearest neighbourhood used in the editing stage to select an appropriate subset of those k nearest centroid

neighbours. The procedure consists of the following steps:

1. Find the k nearest centroid neighbours of q.
2. Select the $j \leqslant k$ nearest centroid neighbours correctly labelled according to the NN rule.
3. Assign q to the class with a majority of votes among those j nearest centroid neighbours (resolve ties randomly).

Alternatively, one might use the general k-NN decision rule (instead of the particular 1-NN) for editing the set of the nearest centroid neighbours (Step 2 of the previous algorithm). Nevertheless, in general the use of the k-NN scheme in the editing procedures does not lead to a significantly better behaviour than the simple 1-NN rule.

4. Empirical comparison

Several experiments using both synthetic and real databases (Murphy and Aha, 1991) have been carried out in order to compare the efficiency of the classification schemes considered in this work. Five different random partitions (half of prototypes for training and half for testing purposes) of each original data set, have been used to obtain averaged measures about the performance of each classification rule. In particular, the focus of the present experimental study is on a comparison of the plain k-NCN rule and its modifications with the standard k-NN approach.

4.1. The synthetic databases

This experiment consists of seven data sets corresponding to the same problem but with dimensionality ranging from 2 to 8. These are two classes consisting of multivariate normal distributions with zero mean and standard deviation 1 and 2 in all dimensions, respectively. There are 2500 patterns available in each class.The purpose of this experiment is the study of the behaviour of the classification rules for different dimensionalities in a problem with strong overlap among classes. In this experiment, the results correspond to the best values of k from all the neighbourhood sizes tried.

All classification schemes gave quite similar performances up to dimension 4 (Fig. 1). Nevertheless, there is a clear separation in their behaviour as the dimensionality increases above 5: the k-NCN rule and its modifications obtain much better results than the classical k-NN. Note that the best overall classification rates slightly correspond to the edited k-NCN approach, followed by those of the averaged k-NCN rule. It appears from these results that the k-NCN and specially its modifications are less sensitive to the ratio of size to intrinsic dimensionality of the training set.

Table 1 provides the best values of the parameter k in each dimension. As can be seen, the differences with respect to neighbourhood sizes are not significant. Only for the higher dimensionalities, both modifications of the k-NCN rule need a larger number of neighbours than the plain approach. Nevertheless, it is worth mentioning that the correct classification rates achieved by both

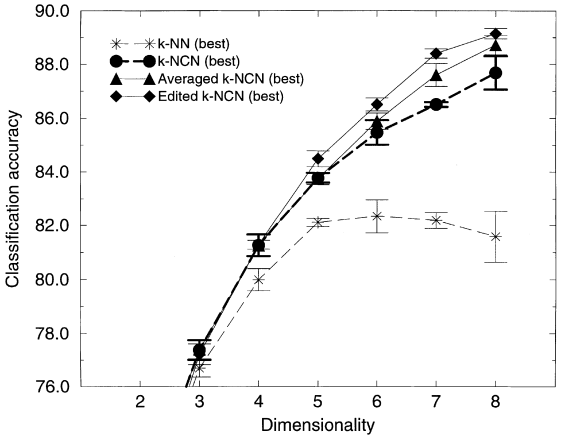

Fig. 1. The synthetic database: classification accuracy with varying dimensionality.

Table 1
The best values of k in each dimension

Dim	Plain k-NN	Plain k-NCN	Averaged k-NCN	Edited k-NCN
3	11	11	11	11
4	11	11	11	11
5	11	11	11	11
6	7	11	11	11
7	5	7	11	9
8	3	5	9	9

extensions with a lower neighbourhood size are very similar to those of the best values of k.

4.2. The texture database

This second database was generated from the Brodatz's photographic album to study a texture discrimination problem with high order statistics. The aim is to distinguish among 11 different textures, each pattern (pixel) being characterized by 40 attributes built by the estimation of fourth order modified moments with four distinct orientations. There are 5500 prototypes, 500 per class.

Fig. 2 illustrates the correct classification rates achieved by the different schemes in this particular problem. Even though all classifiers exhibit quite similar performances, it can be seen that both the plain k-NCN decision rule and its modifications obtain the best accuracy levels, which are significantly better than those of the conventional k-NN approach. In this case, the averaged k-NCN rule provides the highest correct classification rate, although the better result of the edited approach is indeed very close to it.

4.3. The Landsat image database

The purpose of the third experiment is the classification of the multi-spectral values of a real image (2340 × 3380 pixels) taken from the Landsat satellite. This database results from a sub-area of an image, consisting of 82 × 100 pixels. There is a total of 6435 samples (each one corresponding to a 3 × 3 square neighbourhood of pixels completely contained within that sub-area) with 36 characteristics (4 spectral bands × 9 pixels in neighbourhood) and six distinct classes.

The results reported in Fig. 3 are similar to those corresponding to the previous databases. Thus, the k-NN rule gives significantly worse results than the k-NCN schemes. The edited modification increases further the performance, so does the averaged k-NCN rule. Nevertheless, with respect to both alternatives, it seems difficult to make a final statement about the suitability of either of them.

5. Conclusions

Alternative approaches to neighbourhood-based classification have been considered in this work. In particular, the recently introduced k-NCN decision rule along with two heuristic modifications have been used. These extensions to the k-NCN technique try to take into account both proximity and geometrical distribution of the prototypes.

From the experiments carried out, it can be concluded that the modifications proposed here achieve even higher classification rates than the plain k-NCN rule, which has already been shown

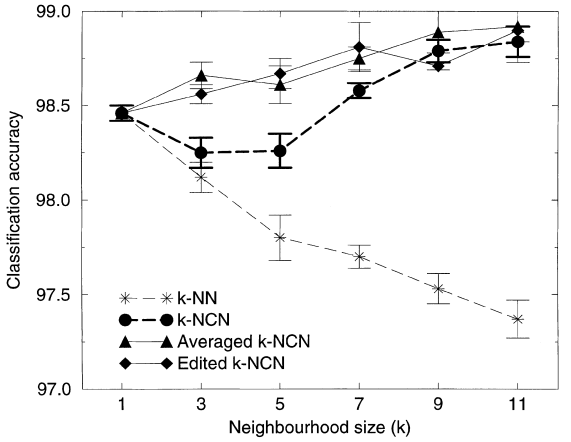

Fig. 2. The texture database: classification accuracy with various neighbourhood sizes.

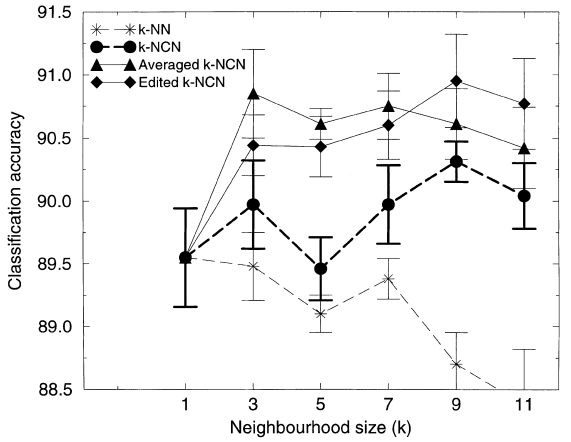

Fig. 3. The Landsat image database: classification accuracy with various neighbourhood sizes.

to give significantly better results than the classical k-NN approach (Sánchez et al., 1997). Nevertheless, although both alternatives increase further the classification performance of the k-NN rule, it seems that the major improvement is mainly due to the properties of the k-NCN technique.

Other modifications of the k-NCN decision rule have been tried but they provided similar or poorer results than the plain classifier. In particular, we have used a pseudo-product classification scheme with the k-NN and k-NCN rules. This approach basically consists of classifying a test sample from its *nearest surrounding neighbours*, that is, those nearest centroid neighbours which are also the nearest neighbours to the given sample.

References

Chaudhuri, B.B., 1996. A new definition of neighbourhood of a point in multi-dimensional space. Pattern Recognition Letters 17, 11–17.

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

Devijver, P.A., Kittler, J., 1982. Pattern Recognition: A statistical Approach. Prentice-Hall, Englewood Cliffs, NJ.

Duda, R., Hart, P.E., 1973. Pattern Classification and Scene Analysis. Wiley, New York.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition. Academic Press, San Diego, CA.

Jaromczyk, J.W., Toussaint, G.T., 1992. Relative neighbourhood graphs and their relatives. Proc. of IEEE 80, 1502–1517.

Murphy, P.M., Aha, D.W., 1991. UCI repository of machine learning databases. University of California, Department of Information and Computer Science, Irvine.

Rabiner, L.R., Levinson, S.E., Rosenberg, A.E., Wilson, J., 1979. Speaker-independent recognition of isolated words using clustering techniques. IEEE Trans. on Acoustics, Speech and Signal Processing 27, 339–349.

Short, R.D., Fukunaga, K., 1980. A new nearest neighbour distance measure. In: Proceedings of the Fifth International Conference on Pattern Recognition. pp. 81–86.

Sánchez, J.S., Pla, F., Ferri, F.J., 1997. On the use of neighbourhood-based non-parametric classifiers. Pattern Recognition Letters 18, 1179–1186.