ELSEVIER

Pattern Recognition Letters 19 (1998) 57–62

Pattern Recognition Letters

A parallel network of modified 1-NN and *k*-NN classifiers – application to remote-sensing image classification

Adam Jóźwik [a,*], Sebastiano Serpico [b], Fabio Roli [b]

[a] *Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, Trojdena 4, 02-109 Warsaw, Poland*
[b] *D.I.B.E. Department of Biophysical and Electronic Engineering, University of Genova, Via Opera Pia 11A, 16145 Genova, Italy*

Received 9 January 1997; revised 22 September 1997

Abstract

A parallel network of modified 1-NN classifiers and *k*-NN classifiers is described and compared with a standard *k*-NN classifier. All the component classifiers decide between two classes only. The number of all possible pairs of classes determines the number of the component classifiers. The global decision is formed by voting of all the component classifiers. Each of the component classifiers operates as follows. For each class i a certain area A_i is constructed in such a way that area A_i covers all training samples from the class i and possibly a small number of training samples from other classes. In the classification phase, if a sample lies outside of all areas A_i, then the classification is refused. When it belongs only to one of the areas A_i, then the classification is performed by the 1-NN rule. Samples that lie in an overlapping area of some A_i are classified by the *k*-NN rule. Such a classification rule, in this paper called a combined (1-NN, *k*-NN) rule, is used by all component classifiers. Two feature selection sessions are recommended for each of the component classifiers: one to minimize the size of the overlapping areas and another to minimize the error rate for the *k*-NN rule. The aim of this work is to create a classifier with improved performance compared to the standard *k*-NN rule. It is shown that the replacement of the *k*-NN rule by the combined (1-NN, *k*-NN) rule reduces computing time required for classification while the parallelization of the classifier structure decreases the error rate. The effectiveness of the proposed approach was verified on a real data set of 5 classes, 15 features and 8839 samples which was derived from a couple of multisensorial remote-sensing images. © 1998 Elsevier Science B.V.

Keywords: Statistical pattern recognition; Nonparametric method; *k*-NN rule; Parallel classifier

1. Problem statement

The *k*-NN rules belong to the most effective classification methods. Various rules of this type have been defined. The first ones were the 1-NN and the *k*-NN rules analysed by Fix and Hodges (1952); other examples are the edited (k,k')-NN rule considered by Hellman (1970) and the fuzzy *k*-NN rule defined by Jóźwik (1983). We consider now only the standard *k*-NN rule and a net of two-decision classifiers based on this rule. However, our considerations may be valid also for other classifiers.

Usually, the reference sets, i.e. the training sets, contain redundant features and feature selection can significantly improve performances. Feature selection requires a review of some feature combinations. An error rate must be calculated for each reviewed

* Corresponding author.

PII S0167-8655(97)00155-4

feature combination. We use the 'leaving-one-out' method, proposed by Lachenbruch (1965), for this purpose. For the *k*-NN rule and the training set with *m* samples, just as many computations are required as for the 'test set' method with $m-1$ samples in the reference set and *m* samples in the test set. In each of these two cases, we need to find the *k* nearest neighbours out of $m-1$ samples and in both cases this task must be performed *m* times.

The feature selection for the *k*-NN rule needs to determine the optimum number *k* for each reviewed feature combination. It is obvious that the comparison of any two feature combinations for the same fixed value of *k* would be incorrect. To find the optimum value of *k* we estimate the error rates er_k for $k = 1,2,\ldots,m-1$ and select the value that offers the smallest misclassification rate. The determination of *k* for large training sets may require too many computations to be applied. For this reason, if the reference set is large, then we replace the full feature combination review by the well-known forward and backward feature selection strategies described by Devijver and Kittler (1982).

Below, we propose a combined (1-NN, *k*-NN) rule to save the speed as for 1-NN rule and the performance as for the *k*-NN rule. The combined (1-NN, *k*-NN) rule was first suggested by Jóźwik et al. (1993) and recently presented in greater detail by Jóźwik et al. (1996). Next, we propose a parallel net of such two-decision classifiers. The idea of the parallel structure, built with the two-decision *k*-NN classifiers, was introduced eight years ago by Jóźwik and Vernazza (1988) and analysed also by Jóźwik (1994). In this paper, we consider the parallel net of the combined (1-NN, *k*-NN) two-decision classifiers, proposed for the first time by Jóźwik et al. (1994), and then we compare the performance of this net with the standard *k*-NN classifier.

2. Modified 1-NN rule

Let us assume that *nc* is the number of classes and that the reference set consists of *nc* subsets: $X_1, X_2, \ldots, X_{nc}$ and each X_i contains samples from the class *i* only. We associate these sets to the positive real numbers $e_1, e_2, \ldots, e_{nc}$ defined as follows:

$$e_i = \max_{x_j \in X_i} d(X_i - x_j, x_j), \tag{1}$$

where *d* denotes a distance function between a sample x_j and a set $X_i - x_j$ and x_j is an element of X_i, i.e. *d* is the distance between x_j and the nearest sample in $X_i - x_j$. We also define the areas $A_1, A_2, \ldots, A_{nc}$:

$$A_i = \{x : d(X_i, x) \leqslant e_i\}. \tag{2}$$

The areas A_i for the Euclidean distance measure are illustrated in Fig. 1.

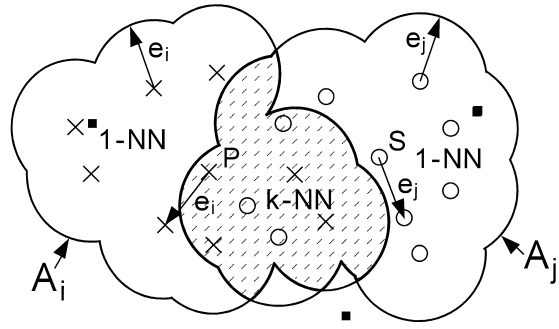

Fig. 1. The combined (1-NN, *k*-NN) classifier for two-class problems. Points with farthest nearest neighbours from the same class, for the classes *i* and *j*, are marked as P and S, respectively.

Now, we can formulate the modified 1-NN rule. A sample represented in the feature space by a point *x* is assigned to the class *i* if it belongs to the area A_i and does not belong to any other A_j, where *j* differs from *i*. If *x* lies outside of each A_i, $i = 1,2,\ldots,nc$, then the classification is refused. When *x* belongs simultaneously to more than one area A_i, then the classifier selects the class which is most heavily represented in the intersection of these A_i. Possible ties can be broken by choosing the largest class in the reference set.

Let us denote by *A* the set of all points in the feature space each of which belongs only to one of the areas A_i, $i = 1,2,\ldots,nc$. It is easy to notice that each sample of the training set that does not belong simultaneously to two or more areas A_i is an element of *A*. Feature selection for the proposed 1-NN rule as modified above consists in finding a feature subset that maximizes the number of samples of the training set in the set *A*.

3. Combined (1-NN, *k*-NN) rule

The modified 1-NN rule can assign the class membership or refuse the classification. This rule can be improved if the samples which lie in the intersections of some A_i are classified by the *k*-NN rule with *k* nearest neighbours found in the whole reference set. Thus, some of these *k* nearest neighbours can lie outside of the intersection of all A_i containing the classified sample. In this manner we have created the combined (1-NN, *k*-NN) rule. So, if the classified sample belongs exactly to one area A_i, then the decision is based on the 1-NN rule; if it lies in the intersection of some areas A_i, then the decision is derived by the *k*-NN rule. When the classified sample lies outside of each A_i, then the classification is refused. Thus, the 'easy' samples can be recognized by the simple 1-NN rule, while the 'difficult' ones are recognized by the more sophisticated *k*-NN rule. The combined (1-NN, *k*-NN) classifier for two classes is presented in Fig. 1.

To find the optimum value of k we need to find all error rates er_k, for $k = 1,2,\ldots,m-1$, where m is the number of samples in the reference set. This can be done by the application of the leave-one-out method, for instance. The values of er_k refer to the whole classifier (not only the *k*-NN classifier) which uses the combined (1-NN, *k*-NN) rule. Some of the samples are classified by the 1-NN rule while others by the *k*-NN rule. The sample that belongs exactly to one area A_i cannot be misclassified while performing the leaving-one-out method by virtue of definitions in Eqs. (1) and (2). Now, we shall give a more detailed explanation of this fact.

The numbers e_i are always the same when the leaving-one-out method is performed, i.e. they are not calculated each time when the one sample is 'taken out'. However, some of the areas A_i may change each time when one sample $y \in X_i$ is 'taken out'. Then, instead of the Eq. (2), we have

$$A_i = \{x : d(X_i - y, x) \leqslant e_i\}. \tag{3}$$

Eq. (1) guarantees that for each sample from the reference set the classification cannot be refused. If during the realization of the leaving-one-out method the object y has been taken from the set X_i then it lies in the area A_i defined by Eq. (3) by virtue of Eq. (1). If the sample y belongs to A_i and does not belong to any other area A_j, $i \neq j$, then the classification is performed by the 1-NN rule and it is correct since the distance of the sample to the set $X_i - y$ is smaller or equal to e_i while its distance to any other X_j, $i \neq j$, is greater than e_i. Thus, only the samples from the intersections of some A_i can be misclassified by the combined (1-NN, *k*-NN) rule. Hence, the only errors done by the combined (1-NN, *k*-NN) classifier during the leaving-one-out are those made by use of the *k*-NN rule.

The above considerations show that the calculations can be restricted to the samples which lie in the intersections of some areas A_i. This means that only samples from the intersections of some A_i are classified when the leave-one-out method is performed.

Two feature selection sessions are recommended. One for the modified 1-NN rule to maximise the number of samples from the training set in the previously mentioned set A and another to minimise an error rate for the *k*-NN rule. The expected advantage of the proposed modification is a significant acceleration of the search for the optimum value of k as well as of the classification itself.

4. Parallel net of combined (1-NN, *k*-NN) classifiers

A multi-class problem can be reduced to some two-decision tasks. One of the possible solutions may be the construction of a parallel net of two-decision classifiers, a separate classifier for each pair of classes, and then forming the final decision by voting of these two-decision classifiers. The illustration of the proposed network, for a 5 class case, is shown in Fig. 2. We shall consider such a network with the component classifiers based on the combined two-

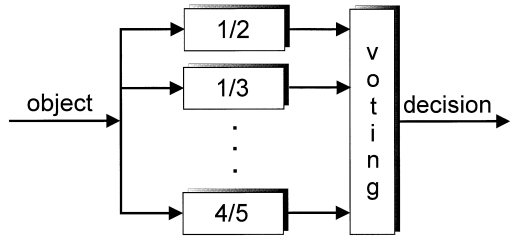

Fig. 2. The parallel net of the two-decision combined (1-NN, *k*-NN) classifiers.

decision (1-NN, k-NN) rules. The component classifiers which refuse the decision (this happens when the classified sample lies outside of all areas A_i) do not take part in voting.

This parallel network of two-decision classifiers should offer better performances than the simple combined (1-NN, k-NN) classifier or the standard k-NN rule. It derives from the geometrical interpretation of both the discussed types of classifiers. In the case of the standard classifier, the boundary separating any pair of classes i and j depends also on the samples from the remaining classes. They have an influence on the value k and on the selected features. These samples may act as a noise. The parallel net may reduce this noise effect. By using the error rate estimated by the leaving-one-out method as a criterion, we can find an optimum number of k for the k-NN rules and perform the feature selection separately for each of the component classifiers.

5. Description of the analysed data set

The algorithm presented above was applied to multisensorial remote-sensing images. In particular, we considered images acquired by two sensors installed on an aircraft: a Daedalus 1268 Airborne Thematic Mapper (ATM) scanner, and a PLC-band, fully polarimetric, NASA/JPL SAR sensor. The flights took place in July and August 1989, respectively. The geographical location was near the Feltwell village (UK), in an agricultural area. More detailed information concerning these images can be found in the paper by Serpico and Roli (1995).

A section (250×350 pixels) of ATM image was selected and registered on the SAR image with an average registration error of about 1 pixel. Then, we extracted 8839 samples (i.e. pixels) belonging to the following classes of agricultural fields: carrots, potatoes, stubble, sugar beets, and wheat. For each sample, a set of 15 features was computed, including original optical and radar channels, texture features calculated by using the above channels and combinations of optical bands.

The following is a brief feature description. Features from 1 to 6 are the six responses of the optical sensor (i.e. the Daedalus sensor) for the bands corresponding to the six channels of the TM sensor in the visible and in the infrared spectrum (the thermal band was disregarded). Features 7 to 9 are responses of the SAR sensor in three channels: band C, polarisation HH (feature 7); band L, polarisation HV (feature 8); and band P, polarisation VV (feature 9). Features 10 to 12 are texture features computed from three channels of the SAR sensor: C band, polarisation HH (feature 10); band C, polarisation HV (feature 11); and band P, polarisation HV (feature 12). Features 13 to 15 are synthetic features computed from the optical sensor as ratios between the responses in different bands: the response in band 7 divided by the sum of the responses in bands 5, 7 and 9 (feature 13); the response in band 9 divided by the sum of the responses in bands 5, 7 and 9 (feature 14); and the response in band 7 divided by the sum of the responses in bands 3, 5 and 7 (feature 15). As a texture feature, we used the fractal dimension of the image signal, considered as a surface, in a neighbourhood of each sample (Peleg et al., 1984).

6. Results

The entire data set contained 5 classes, 15 features and 8839 samples, as described above. Ten different experiments were carried out, with 2440 samples chosen at random for the training set while the remaining 6399 samples were used for the test set. As a distance function the Euclidean measure was taken.

We started our computations with the modified 1-NN rule for the entire training set by considering all 5 classes together, though unsuccessfully. Only five samples from the training set were found in the set A defined in Section 3, in the first experiment, although the feature selection was performed to maximise the size of A. The two feature selection strategies 'forward' and 'backward' were applied in each of the 10 cases and the best of these two results was finally taken. In the remaining 9 experiments the situation was similar, between 3 to 7 samples in A. Without feature selection the results were even worse. In such a situation, when nearly all samples lie in the intersection of some areas A_i, the use of the combined (1-NN, k-NN) rule instead of the standard k-NN classifier does not offer any advantage.

Table 1
Feature selection results and overlap rates (r/m_{ij}) for the first experiment using the modified 1-NN rule

Pair of classes	Selected features	Overlap rate in %
1 and 2	2,3,4,6	8.1
1 and 3	4,5,6	6.0
1 and 4	2,3,4,6	7.5
1 and 5	2,3,4,6	8.1
2 and 3	4,6	2.5
2 and 4	4,5,6	6.1
2 and 5	4,6	2.7
3 and 4	4,6	2.1
3 and 5	1,3,4,5,7,9,12,13	99.8
4 and 5	4,6	6.0

Next, we performed the computations for the parallel net of two-decision (1-NN, k-NN) classifiers as proposed above. The feature selection strategies were the same as in the previous unsuccessful experiment. Both of them were always applied and the best result finally taken.

The feature selection results for the first of the 10 experiments and the modified 1-NN rule are given in Table 1. The overlap rate means the ratio r/m_{ij}, where r is the number of samples found in the intersection of the areas A_i and A_j and m_{ij} is the number of samples in the training set that corresponds to the pair of classes i and j.

Feature selection results for the k-NN rule, restricted to the overlap areas, are presented in Table 2.

The optimum values of k for the component classifiers found by the leave-one-out method varied between 1 and 7. The final result obtained by the parallel network of combined (1-NN, k-NN) two-decision classifiers and calculated with the use of the test set was $e = 0.0135$, i.e., below 1.5%.

As we can see in Table 1 (for the pair of classes 3 and 5) the proposed modification of the 1-NN rule can sometimes fail. For this pair of classes the standard k-NN rule and the combined (1-NN, k-NN) offered exactly the same result as that given in Table 2.

The results given in Table 1 show that the determination of the optimum k and the classification by the k-NN rule can be restricted to a very small number of samples (except for the pair 3 and 5).

Tables 1 and 2 concern only the first one of the ten experiments.

Table 3, given below, presents results which can be used to compare the performance of the parallel net of the combined (1-NN, k-NN) classifiers with the standard k-NN classifiers. All error rates were calculated with the use of the test sets. The first two columns show results for the simple k-NN rules (S k-NN) with all features (no feature selection) and with the selected features (with f.s.), respectively. The next two columns concern the proposed parallel net of two-decision classifiers (P k-NN). Feature

Table 2
Feature selection results and error rates for the k-NN rule, restricted to the overlap areas

Pair of classes	Selected features	Error rate in %
1 and 2	2,4,5	2.1
1 and 3	1,4,5,6,9,11	0.7
1 and 4	1,3,5,6,12,15	3.2
1 and 5	3,5,8,11,14,15	2.4
2 and 3	2,6,8,10,15	0.8
2 and 4	1,4,6,11,12	1.1
2 and 5	3,5,6,8,12	0.8
3 and 4	1,6,8	0.0
3 and 5	2,4,6,9	0.0
4 and 5	2,6,8,10,15	0.8

Table 3
Comparison of error rates of the simple k-NN rules (S k-NN) with all features (no feature selection) and with the selected features (with f.s.) and the proposed parallel net of two-decision classifiers (P k-NN)

Exp. no.	S k-NN		P k-NN	
	no f.s.	with f.s.	no f.s.	with f.s.
1	2.01	1.64	1.64	1.35
2	2.25	1.80	1.93	1.35
3	1.68	1.52	1.56	1.02
4	1.76	1.48	1.43	1.35
5	2.01	1.64	1.43	1.27
6	1.56	1.27	1.39	1.31
7	1.84	1.48	1.60	1.27
8	2.09	1.68	1.76	1.31
9	2.21	2.05	1.84	1.93
10	2.25	1.84	1.80	1.60
Minimum	1.56	1.27	1.39	1.02
Maximum	2.25	2.05	1.93	1.93
Average	1.97	1.64	1.64	1.38

selection was applied to each of the component classifiers, separately for the modified 1-NN classifiers and *k*-NN classifiers which realise the combined (1-NN, *k*-NN) rules.

7. Concluding remarks

The amount of computations required to estimate an error rate for the *k*-NN rule grows rapidly as the size of the reference set gets larger. One way to reduce the number of these computations is to break down the large data classification problem into some smaller tasks. Hence, instead of one *k*-NN multi-decision classifier we could apply the parallel net of two-decision classifiers. Another way to reduce the number of computations is to restrict an application of the sophisticated methods to 'difficult' samples only.

The considerations presented above show that the second approach may fail. However, it does not involve a lot of computations and, for this reason, is worth checking. The first way, i.e. the network solution, is universal by nature and can always be applied if the number of classes is bigger than two.

As we can notice from Tables 1 and 2, the first stage required smaller numbers of features. This is a desirable phenomenon since it concerns the larger data sets.

Table 1 indicates that the number of samples that must be sent to the *k*-NN classifier was reduced, in 9 of 10 cases, more than 10 times.

The proposed net, as shown in Table 3, offers remarkably better performance than the standard *k*-NN rules. The standard *k*-NN classifiers, after the feature selection, required most often 11 features while the component (1-NN, *k*-NN) classifiers needed usually no more than 4 and 7 features for the 1-NN and the *k*-NN rule, respectively. However, the total number of features required by the proposed net was always equal to 15, i.e., all features were used.

It would be interesting to investigate how the proposed net and the standard *k*-NN classifier would behave as the sizes of the training sets grow, starting, for instance, with 100 samples and stopping at 10000 samples. It seems that the parallel net would converge faster to the Bayes classifier. It means that for smaller training sets the proposed net would be even more competitive. This problem will be a subject of our further studies.

Acknowledgements

This research has been supported by the COPERNICUS programme within the project CRAck and SHape Defect Detection in Ferrite Cores (CRASH), contract: No. CIPA-CT94-0153. The part of the results presented has been obtained by the Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, Poland.

References

Devijver, P.A., Kittler, J., 1982. Pattern Recognition: A Statistical Approach. Prentice Hall International, London.

Fix, E., Hodges, J.L., 1952. Disriminatory analysis: Nonparametric discrimination: Small sample performance. Project 21-49-004, Report Number 11, USAF School of Aviation Medicine, Randolph Field, TX, pp. 280–322 (Reprinted in the book: Dasarathy, B.V., 1991. NN Pattern Classification Techniques. IEEE Computer Society Press, Silver Spring, MD, pp. 40–56).

Hellman, M.E., 1970. The Nearest Neighbor classification rule with a reject option. IEEE Trans. System Sci. Cybernet. 6 (3), 179–185.

Jóźwik, A., 1983. A learning scheme for a fuzzy *k*-NN rule. Pattern Recognition Letters 1 (5/6), 287–289.

Jóźwik, A., 1994. Pattern recognition method based on *k* Nearest Neighbor rule. J. Communications 45, 27–29.

Jóźwik, A., Vernazza, G., 1988. Recognition of leucocytes by parallel *k*-NN classifiers. Lecture Notes of ICB Seminar, Warsaw, pp. 138–153.

Jóźwik, A., Roli, F., Dambra, C., 1993. A multistage synthesis of modified NN rule and its application for remote-sensing images. In: Proc. IPTA 1993, San Remo, Italy, pp. 435–438.

Jóźwik, A., Serpico, S., Roli, F., 1994. Network of modified 1-NN and fuzzy *k*-NN classifiers in application to remote sensing image recognition. In: EUROPTO Series, Proceedings, Vol. 2315, Rome, pp. 26–30.

Jóźwik, A., Chmielewski, L., Cudny, W., Skłodowski, M., 1996. A 1-NN preclassifier for fuzzy *k*-NN rule. In: Proc. 13th Internat. Conf. on Pattern Recognition, Vol. 4, Track D, Vienna, pp. 234–238.

Lachenbruch, P.A., 1965. Estimation of error rates in discriminant analysis. Ph.D. Dissertation, University of California, Los Angeles, Chapter 5.

Peleg, S., Naor, J., Hartley, R., Avnir, D., 1984. Multiple resolution texture analysis and classification. IEEE Trans. Pattern Anal. Machine Intell. 6 (4), 518–523.

Serpico, S.B., Roli, F., 1995. Classification of multisensor remote sensing images by structured neural networks. IEEE Trans. Geoscience and Remote Sensing 33 (3), 562–578.