Adaptive Selection of Image Classifiers

Giorgio Giacinto and Fabio Roli

Dept. of Electrical and Electronic Eng., University of Cagliari, Italy
Piazza D'Armi, 09123, Cagliari, Italy - Phone: +39-70-6755874 Fax: +39-70-6755900
e-mail {giacinto,roli}@diee.unica.it

Abstract. Recently, the concept of "Multiple Classifier Systems" was proposed as a new approach to the development of high performance image classification systems. Multiple Classifier Systems can be used to improve classification accuracy by combining the outputs of classifiers making "uncorrelated" errors. Unfortunately, in real image recognition problems, it may be very difficult to design an ensemble of classifiers that satisfies this assumption. In this paper, we propose a different approach based on the concept of "adaptive selection" of multiple classifiers in order to select the most appropriate classifier for each input pattern. We point out that adaptive selection does not require the assumption of uncorrelated errors, thus simplifying the choice of classifiers forming a Multiple Classifier System. Reported results on the classification of remote-sensing images show that adaptive selection can be used to obtain substantial improvements in classification accuracy.

1 Introduction

Recently, in the field of character recognition, the concept of Multiple Classifier Systems (MCSs) was proposed as an approach to develop a high performance recognition system [1,2]. In particular, it has been pointed out that by combining the outputs of an MCS it is easy to exploit complementary characteristics of classification algorithms based on different methodologies and/or using different input features [1]. The potentialities of these recognition systems have been reported also in the remote sensing field [3–5]. Several combination functions have been proposed based on voting rules, statistical theory, Dempster-Shafer evidence theory, belief functions, and many other "integration schemes" [1,2,6–8]. Despite the promising results reported in the literature, performances of MCS greatly depend on the assumption that classifiers exhibit a sufficiently large "uncorrelation" in their classification errors [1,9,10].

In this paper, a different approach to the exploitation of the potential advantages of MCSs is proposed. This approach is based on the concept of "adaptive selection" of multiple classifiers aimed at selecting the most appropriate classifier for each input pattern. This concept is not completely new in the field of pattern recognition. Recently, Srihari et al. pointed out the potentialities of "dynamic classifier selection" [7]. In the neural networks field, Jacobs and Jordan proposed a multiple neural network system that allows for a particular kind of

dynamic selection based on the concept of "adaptive mixtures of local experts" [11]. In this paper, we first point out that adaptive selection does not require the assumption of uncorrelated errors (Section 2). Afterwards, a selection algorithm is described (Section 3). In Section 4, a method based on data clustering is proposed to design selection-based multiple classifier systems. In Section 5, experimental results on the classification of remote-sensing images are reported. Conclusions are drawn in Section 6.

2 Multiple Classifier Systems: Selection vs. Combination

Some researchers clearly showed that combination mechanisms can increase classification accuracy only if the assumption of independent classification errors is satisfied. Hansen and Salamon showed that a combination mechanism based on a simple majority decision rule can provide very good performances if classifiers are "independent" [8]. Tumer and Gosh pointed out that classification accuracy increases obtained by combining depend on error correlation more than on the particular combination mechanism adopted [9]. On the other hand, experimental results showed that, in real pattern recognition applications, may be very difficult to design and train independent classifiers, even if based on different methodologies [2,3]. Consequently, very recently, some researchers proposed combination mechanisms aimed at avoiding the independence assumption [4,12]. Methods that identify and remove classifiers that are excessively correlated have also been proposed [10,13].

It is quite easy to see that if we could design an "optimal classifier selector" that always selects the most appropriate classifier for each test pattern, then there would be no longer any need for an ensemble of independent classifiers. For each test pattern, it would be sufficient to have just one classifier that correctly classifies it. Unfortunately, it is just as easy to see that the above optimal classifier selector is more difficult to "design" than the combination mechanisms adopted in the present MCSs. The design of an adaptive classifier selector requires the definition of "selecting conditions" that focuses on choosing the most appropriate classifier for each input pattern. On the other hand, the combination can be implemented more simply, but requires the "selection" of independent classifiers. Therefore, selection mechanisms can greatly simplify that part of the MCS design related to a choice of classifiers. Their drawbacks are mainly related to designing complexity and computational load. The reverse is true for combination mechanisms.

3 The Proposed Algorithm for Adaptive Classifier Selection

The proposed algorithm is based on the definition of a "selecting condition" which makes it possible to select, for each test pattern, the classifier that has more chances to make a correct classification on that pattern. This selecting

condition is based on the estimate of *classifier local accuracies* in a "neighbourhood" of the input pattern $\mathbf{X}$ (neighbourhood($\mathbf{X}$)), defined with respect to a "validation set", i.e., a set of data whose classification is known but that is different from the data set used to train classifiers. The neighbourhood could be also defined with respect to the training set, but it may be very difficult to provide good estimates of classification local accuracies, since, mainly due to the so-called "overfitting problem", classifiers exhibit good performances.

Let us assume that our MCS is formed by K classifiers C_j, $j = 1 \ldots K$ and each classifier focuses on solving a pattern recognition problem with M data classes ω_i, $i = 1 \ldots M$. For each test pattern $\mathbf{X}$ the estimate of classifier local accuracies in a "neighbourhood" of the input pattern $\mathbf{X}$ can be computed with the following formula:

$$\hat{p}(correct_j/\mathbf{X}, neighbourhood(\mathbf{X})) = \frac{N_j}{N} \tag{1}$$

where N is the number of validation patterns forming the neighbourhood($\mathbf{X}$) and N_j is the number of validation patterns that were correctly classified by the classifier C_j. At present, the appropriate dimension of the neighbourhood is decided by experiments or by using heuristic rules.

The ratio computed in the above equation is assumed to be equal to the probability that classifier C_j correctly classifies the test pattern $\mathbf{X}$. The rationale of this assumption is a sort of "stationarity" of classification accuracy in a small "partition" of the data set, i.e., all the patterns belonging to the neighbourhood have the same probability of being correctly classified by a given classifier. The general validity of this assumption is very difficult to prove. It strictly depends on the available data set and on the size of the neighbourhood. However, according to our experiments, it seems to apply for most cases and, in particular, it is reasonable for our purposes, since it allows us to compare the classifiers "locally" in order to select the most appropriate for the test pattern.

A "soft" version of equation (1) can be defined as follows:

$$\hat{p}(correct_j/\mathbf{X}, neighbourhood(\mathbf{X})) = \frac{\sum_{i=1}^{N} p_j(\omega_k/\mathbf{X}_i \in \omega_k) W_i}{\sum_{i=1}^{N} W_i} \tag{2}$$

where:

- ω_k ($k = 1 \ldots M$) is the correct data class for the neighbourhood pattern $\mathbf{X}_i$;
- $p_j(\omega_k/\mathbf{X}_i)$ is an estimate of the posterior probability provided by classifier C_j. This term constitutes a measure of classifier accuracy on the validation pattern $\mathbf{X}_i$ and, with respect to the "hard" selecting condition defined by equation (1), allows uncertainties related to validation data classifications to be managed more efficiently;
- $W_i = 1/d_i$, where d_i is the Euclidean distance of validation pattern $\mathbf{X}_i$ from the test pattern $\mathbf{X}$. This term takes into account the uncertainty due to the heuristic neighbourhood-size definition.

It is easy to see that equations (1) and (2) have a value equal to 1 when the classifier C_j perfectly classifies all the neighbourhood patterns.

The following adaptive classifiers selection algorithm was defined on the basis of the selection conditions described above. Equations (1) or (2) can be used to implement a "hard" or a "soft" selecting condition, respectively.

****Adaptive Classifiers Selection Algorithm****
Input parameters: classifier confusion matrices on the validation set and
size of the neighbourhood
Begin
For each test pattern **X**:
Do
$\forall C_j (j = 1 \ldots K)$:
Begin
Do
STEP 1: Compute $\hat{p}(correct_j/\mathbf{X}, neighbourhood(\mathbf{X}))$
STEP 2: If $\hat{p}(correct_j/\mathbf{X}, neighbourhood(\mathbf{X})) < 0.5$ ***Then***
Reject classifier C_j
End
STEP 3: Identify the classifier C_m exhibiting the maximum value of
$p(correct_j/\mathbf{X}, neighbourhood(\mathbf{X})), j = 1 \ldots K', K' \leq K$
STEP 4: For each classifier C_l, $l = 1 \ldots K'$, compute the following difference
$d_l = [\hat{p}(correct_m/\mathbf{X}, neighbourhood(\mathbf{X})) - p(correct_l/\mathbf{X}, neighbourhood(\mathbf{X}))]$
STEP 5: If $\forall l, l = 1 \ldots K', l \neq m, d_l > Threshold$ ***Then***
Select Classifier C_m ***Else***
Randomly Select one of classifiers for which $d_l < Threshold$
End

Steps 1 and 2 focus on selecting K' classifiers ($K' \leq K$) by removing classifiers that have a probability of less than 0.5 to correctly classify the test pattern **X**. The differences computed at Step 4 are used to compute a sort of "confidence" for the selection. If all the differences are higher than an a-priori fixed threshold (e.g., 0.1), then there is reasonable confidence that classifier C_m is the most appropriate for the test pattern. On the other hand, a random selection is carried out between C_m and the classifiers that exhibit values of the selecting condition close to the value exhibited by C_m. In fact, it is not reasonable to directly select the classifier C_m if there are other classifiers exhibiting similar values of the selecting condition.

4 A Method for Designing MCSs

The basic concepts of this method are the subdivision of the training set into "partitions" and the assignment of each partition to a "specialised classifier". Each specialised classifier is dedicated to correctly classify a partition of the data set and it is consequently trained only on that partition. It is easy to see that

the operation mechanism of an MCS based on the above specialised classifiers should be an adaptive selection mechanism.

Let us assume that the training data set Ω is defined by the union of M mutually exclusive data classes ω_k:

$$\Omega = \bigcup_{k=1}^{M} \omega_k \tag{3}$$

Analogously, each data class ω_k can be defined by the union of M_k data clusters $\omega_{k,m}$, by using one of the many clustering algorithms proposed in the literature [14]:

$$\omega_k = \bigcup_{m=1}^{M_k} \omega_{k,m} \tag{4}$$

After clustering, the number M_k of clusters is generally different for each data class. Let us assume that ω_i is the data class with the maximum number of clusters M_i. Our goal is to create "partitions" of the data set that correspond to different classification tasks with the same number of M data classes as the initial task, and to assure that the union of these partitions "covers" the data set Ω. To this end, for each data class ω_k, $k \neq i$, "cloned" clusters $\omega_{k,*}$ are generated by a random choice of "natural" clusters in order to obtain a number of clusters equal to M_i for all classes:

$$\omega_k = \omega_{k,1} \cup \omega_{k,2} \cup \cdots \cup \omega_{k,M_k} \cup \omega_{k,*} \cup \cdots \cup \omega_{k,*}, k = 1 \ldots M, k \neq i \tag{5}$$

As an example, if the class ω_1 has two clusters and M_i is equal to four, then two new clusters for the class ω_1 are generated by randomly choosing among the two natural clusters $\omega_{1,1}$ and $\omega_{1,2}$.

Afterwards M_i partitions of the data set P_z, $z = 1 \ldots M_i$ are defined as follows:

$$P_z = \bigcup_{j=1}^{M} \omega_{j,z} \tag{6}$$

and a specific classifier C_z is trained on each partition. In most cases, the above partitions are not mutually exclusive. Therefore, the resulting data set covering is redundant.

5 Experimental Results

5.1 Data Set Description

The data set used for our experiments consists of a set of multisensor remote-sensing images related to an agricultural area near the village of Feltwell (UK) [15]. The images were acquired by an ATM sensor with eleven bands and a SAR with twelve channels, both installed on an airplane. For our experiments each

pixel was characterised by a fifteen-element "feature vector", using six bands of the ATM and nine channels of the SAR. We selected 10944 pixels belonging to five agricultural classes (i.e., sugar beets, stubble, bare soil, potatoes, carrots) and subdivided them into a training set (5124 pixels), a validation set (582 pixels), and a test set (5238 pixels). We used a very small validation set to simulate real cases where validation data are difficult to obtain.

5.2 Results and Comparisons

Several experiments have been carried out to validate the proposed methods [16]. In the following, for the sake of brevity, we report two main experiments (here called Experiments A and B) that show the main advantages provided by the proposed methods.

Experiment A: We designed an MCS consisting of four classifiers: three multilayer perceptrons (MLPs) neural networks with different architectures to make them as "independent" as possible, and one k-nearest neighbour (k-nn) classifier (we used $k = 21$). With regard to the parameters of our selection algorithm, we used a neighbourhood containing twenty validation patterns and the selecting condition was based on equation (2). Table 1 shows classification accuracies on the test set provided by our selection-based MCS compared to those of individual classifiers. The selection-based MCS substantially improves the classification accuracy without increasing the rejection rate. Table 2 shows the comparison between performances of our selection-based MCS and those of MCSs based on two of the most commonly used combination mechanisms proposed in the literature [1], i.e., the "majority rule" and the "Bayesian average". Both of these methods require the assumption of "independent errors". Results show that selection-based MCSs allows one to improve accuracies provided by MCSs based on combination mechanisms. These results agree with our analysis of correlation among errors made by individual classifiers [15,16]. We also compared the *selection performances* provided by our selection mechanism with the "reference" performances provided by a sort of "oracle" that always predicts the best classifier for each test pattern [16]. The proposed selection algorithm was able to make the correct decision on the most appropriate classifier for 97.22% of the test set.

Classification Algorithm	% Accuracy	% Rejection
Selection-based MCS	93.10	1.83
Neural Network 1 MLP 15-30-15-5	87.30	1.66
Neural Network 2 MLP 15-7-7-5	85.36	1.13
Neural Network 3 MLP 15-15-5	90.71	2.83
k-nn Classifier	90.70	1.89

Table 1. Classification accuracies on the test set provided by our selection-based MCS compared to those of individual classifiers

	% Accuracy	% Rejection
Selection-based MCS	93.10	1.83
Majority-based MCS	90.38	3.37
Average-based MCS	89.48	not available

Table 2. Comparison between performances of our selection-based MCS and those of MCSs based on the combination mechanisms

Experiment B: This experiment focused on evaluating performances of MCSs designed according to the method described in Section 4. For this purpose, a clustering algorithm based on a "hierarchical clustering technique" was performed on training data [14]. Different numbers of cluster were found for the five data classes contained in the selected data set (Class 1: 2 clusters, Class 2: 4 clusters, Class 3: 7 clusters, Class 4: 5 clusters, Class 5: 2 clusters). According to the proposed method, the training set was subdivided into seven partitions and a MCS based on seven classifiers was designed. In particular, we used seven k-nearest neighbour classifiers. With regard to the parameters of our selection algorithm, we used a neighbourhood containing six validation patterns and the selecting condition was based on equation (2). Table 3 shows classification performances of MCS designed according to our method. With respect to the performances of an "optimal" selector, the *selection accuracy* obtained by using this method is 95.73%. It is worth noting that these results cannot be directly compared to those contained in the previous Tables, since different classifiers form the related MCSs.

6 Conclusions

In this paper, we proposed a novel approach to the exploitation of potential advantages of MCSs based on the concept of "adaptive selection". We described an "adaptive classifiers selection algorithm" and reported experimental results related to the classification of remote-sensing images. We showed that the proposed selection-based MCS performs better than classical MCSs based on combination mechanisms. In particular, we showed that our selection algorithm provides performances that are reasonably close to those of an optimal selector. Finally, we proposed a systematic method to design MCSs based on classifiers selection and reported the satisfactory classification accuracies provided by MCSs designed according to this method. To the best of our knowledge, no other adaptive classifiers selection algorithm has been presented in the pattern recognition literature.

	% Accuracy	% Rejection
Selection-based MCS	96.30	3.09

Table 3. Performances of MCS designed according to the method in Section 4

In the field of neural networks, only Jordan's work can be regarded as an implementation of the concept of dynamic classifiers selection, since his "mixture of local experts" is adaptive [11]).

References

1. L.Xu, A.Krzyzak, and C.Y.Suen, "Methods for combining multiple classifiers and their applications to handwriting recognition", IEEE Trans. on Systems, Man, and Cyb., Vol. 22, No. 3, May/June 1992, pp. 418-435
2. R.Battiti, and A.M.Colla, "Democracy in neural nets: voting schemes for classification", Neural Networks, Vol. 7, No. 4, 1994, pp. 691-707
3. F.Roli, G.Giacinto, and G.Vernazza, "Comparison and combination of statistical and neural network algorithms for remote-sensing image classification", Neurocomputation in Remote Sensing Data Analysis, Advances in Spatial Science Series, Springer Verlag Ed. (in press, 1997)
4. G.Giacinto, and F.Roli, "Ensembles of Neural Networks for Soft Classification of Remote Sensing Images", Proc. of the European Symposium on Intelligent Techniques, March 20-21, 1997, Bari, Italy, pp.166-170
5. I.Kanellopoulos et al., "Integration of neural network and statistical image classification for land cover mapping", Proc. IGARSS 93, Tokio, 18-21 August 1993, pp. 511-513
6. Y.S.Huang, and C.Y.Suen, "A method of combining multiple experts for the recognition of unconstrained handwritten numerals", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.17, No.1, January 1995, pp.90-94
7. N.Srihari et al., "Decision combination in multiple classifier systems", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.16, No.1, Jan. 1994, pp. 66-75
8. L.K.Hansen, and P.Salamon, "Neural network ensembles", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 12, No. 10, October 1990, pp. 993-1001
9. K.Tumer and J.Gosh, "Error correlation and error reduction in ensemble classifiers", Tech. Report, Dept. of ECE, University of Texas, July 11, 1996
10. D.Partridge, W.B.Yates, "Engineering multiversion neural-net systems", Neural Computation, 8, 1996, pp. 869-893
11. R.Jacobs, M.Jordan, S.Nowlan, and G.Hinton, "Adaptive mixtures of local experts", Neural Computation, 3, 1991, pp. 79-87
12. C.Y.Suen et al., "The combination of multiple classifiers by a neural network approach", Int. Journal of Pattern Recognition and Artificial Intelligence, Vol. 9, no.3, 1995, pp.579-597
13. D.Opitz, and J.Shavlik, "Generating accurate and diverse members of a neural-network ensemble", Advances in Neural Information Processing Systems 8, MIT Press, 1996
14. R.O.Duda, P.E.Hart, "Pattern Classification and Scene Analysis", Wiley & Sons, Inc., 1973
15. S.B.Serpico, L.Bruzzone and F.Roli, "An experimental comparison of neural and statistical non-parametric algorithms for supervised classification of remote-sensing images", Pattern Recognition Letters, Vol 17, No. 13, November 1996, pp. 1331-1341
16. F.Roli and G.Giacinto, "Adaptive selection in multiple classifier systems", Tech. Rep., MCS-4-96, University of Cagliari, Italy, 1996