Classification Reliability and Its Use in Multi-classifier Systems

L. P. Cordella, P. Foggia, C. Sansone, F. Tortorella, M. Vento

Dipartimento di Informatica e Sistemistica
Via Claudio 21, I-80125 Napoli, Italy
E-mail: {cordel, foggia, carlosan, tortorel, vento}@nadis.dis.unina.it
WWW: http://amalfi.dis.unina.it

Abstract

In the last years, great attention has been devoted to multiple classifier systems. The implementation of such a system implies the definition of a rule (combining rule) for determining the most likely class, on the basis of the class attributed by each single expert. The availability of a criterion to evaluate the credibility of the decision taken by a classifier can be profitable in order to implement the combining rule. We propose a method that, after defining the reliability of a classification on the basis of information directly derived from the output of the classifier, uses this information in the context of a combining rule. The results obtained by combining four handwritten character recognizers on the basis of classification reliability are compared with those obtained by using three different combining criteria. Tests have been performed using a standard handwritten character database.

1 Introduction

In many Pattern Recognition applications, achieving acceptable recognition rates is conditioned by the large pattern variability, whose distribution cannot be simply modeled. This affects the results at each stage of the recognition system so that, once this has been designed, its performance cannot be improved over a certain bound, despite the efforts in refining either the classification or the description method.
Employing a multiple classifier system can be very useful for tackling such situation [1,2]: in fact, the consensus of a set of experts may compensate for the weakness of the single expert, while each single expert preserves its own strength. The implementation of a multiple classifier system implies the definition of a combining rule for determining the most likely class a sample should be attributed to, on the basis of the class to which it is attributed by each single expert [2]. The availability of a criterion to evaluate the credibility of the decision taken by a classifier can be very profitable in order to implement the combining rule. However, the definition of a parameter measuring such credibility is quite critical: an improper definition may result in giving high credibility to experts whose classification is unreliable, or viceversa, low credibility to really reliable experts.

Most of the combining rules proposed in the literature implement a "weighted voting", in the sense that the vote (i.e. the attribution to a class) expressed by an expert is weighted by the reliability estimated on the basis of the class chosen [1]: the input sample is then assigned to the class for which the sum of the weighted votes is the highest. In this way, however, all the samples attributed to the same class are assigned

the same reliability and thus this value could not reflect the actual reliability of the single classification act.

To get out of this problem, we propose to associate a reliability measure to each classification performed by a given expert and to use this value to weight its vote in a multi-expert system. The operative definition of the parameter allowing to recognize situations which can give rise to unreliable classifications and enabling to quantify classification reliability will depend on the considered classifier architecture.

To evaluate the effectiveness of the reliability parameter several multi-expert systems, each obtained by combining various experts according to different combining rules, have been employed. The experts are handwritten character recognizers made of different pairs descriptor-classifier which will be described in the following.

Tests have been carried out using the digits of the NIST Database 19 [3]. Note that the experts have not been selected because they perform particularly well on the considered database. Aim of this paper is only that of quantitatively evaluating the differences of performance obtainable when using different definitions for the weights attributed to the votes of the experts in a multi-expert system. In the next Sections the adopted experts are briefly described and the definition of the classification reliability parameter is introduced together with the used combining rules. Finally, the experimental results are presented.

2 The Adopted Experts

In the field of handwritten character recognition character descriptions can be based on measurements directly performed on the character bit map (not structural descriptions) or given in terms of component parts, coming from a decomposition of the character, and relations among them (structural descriptions). Hybrid techniques that combine the two approaches are also possible.

We have considered each of the three types of description schemes. The not structural description uses as features the pixels of the character image obtained after a suitable filtering and scaling process leading to a small size "gray level picture" whose pixel values are computed by averaging the original image and are normalized so as to fall within the interval [0,1]. The obtained 8x8 matrices of numbers, encoded as vectors of length 64, are the descriptors used by the classifiers. Fig. 1 shows some characters before and after the scaling process.

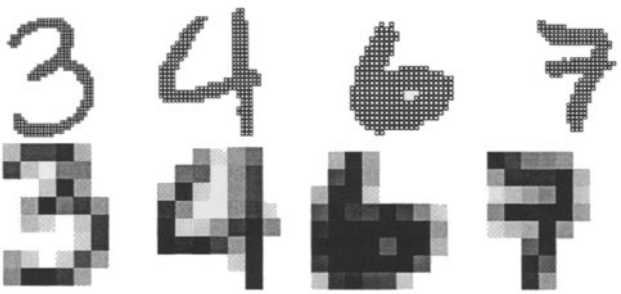

Fig. 1. Some characters of the NIST database and the results of the scaling process (8x8 gray level matrices).

In order to obtain the adopted structural description, characters are thinned (Fig. 2*a*,*b*) and then further processed for correcting the shape distortions introduced by thinning. After this correction a character is represented by a set of polygonal lines (Fig. 2*c*) which are then approximated with pieces of circular arcs (Fig. 2*d*). The structural description can be conveniently put in the form of an Attributed Relational Graph (ARG) (Fig. 2e), whose node attributes specify span, relative size and orientation of the corresponding arc, while branch attributes specify topologic relations between arc pair projections on the coordinate axes. More details can be found in [4].

The description we have called hybrid combines the structural and not structural approaches. After approximating a character with a set of circular arcs (see Fig. 2d), geometrical moments of this set are computed [5]. Geometric moments up to the 7th order are considered. Moments of zero and first order have been used to make the remaining moments invariant with respect to scale and translation. A character is thus described by a 33 element vector.

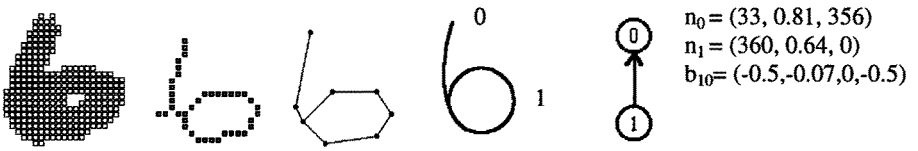

Fig. 2. An example illustrating the structural description process: *a)* Character bit map, *b)* character skeleton, *c)* polygonal approximation of the skeleton, *d)* decomposition in terms of circular arcs, *e)* the corresponding Attributed Relational Graph; attribute values of nodes and branches are listed (n_i represents the i-th node, while b_{ij} denotes the branch connecting the i-th and j-th nodes).

Three classifiers were implemented by two different kinds of neural networks: the Multi-Layer Perceptron (MLP) [6] and the Learning Vector Quantization network (LVQ) [7]; the fourth classifier is of the Nearest Neighbor type (NN).

The above classifiers have been combined with the three previously outlined descriptions, giving rise to the four experts illustrated in the following. The acronyms used to denote the experts specify the classifier and the associated description type. Only handwritten digits were considered for carrying out the test, thus the number of classes is ten for every classifier.

The *MLP-NS Expert:* the MLP-NS expert combines the MLP classifier with the Not Structural description. Therefore the input layer of the classifier is made of 64 neurons each one associated to a pixel of the scaled image. The chosen network architecture has a single hidden layer of 30 neurons and an output layer of 10 neurons corresponding to the ten digits. The learning algorithm is the standard Back-Propagation one, with a constant learning rate equal to 0.5. The sigmoidal activation function was chosen for all the neurons.

The *LVQ-NS Expert:* the LVQ-NS has an input layer composed by 64 neurons and a number of Kohonen neurons fixed to 7 for all classes. The net was trained with a supervised version of the FSCL algorithm to overcome the neuron under-utilization

problem [8]. The learning rate was initially set equal to 0.5 and then decreased according to the rules illustrated in [8].

The *MLP-H Expert:* in this expert, the classifier works with the Hybrid description. Thus the input layer of the classifier is made of 33 neurons. All the remaining network parameters are the same used in the MLP-NS expert.

The *NN-S Expert:* it uses the structural description associated to a NN statistical classifier. The distance between two characters is measured by means of a metric defined in the ARG space [9]. In order to reduce the computational effort, the experiments performed with this expert were carried out using only a subset (about 25%) of the training set used in the other cases.

3 Defining the Reliability Parameters

The low reliability of a classification is generally due to one of the following situations: a) the considered sample is significantly different from those present in the training set, i.e., its representative point is located in a region of the feature space far from those occupied by the samples of the training set and associated to the various classes; b) the point which represents the considered sample in the feature space lies where the regions pertaining to two or more classes overlap, i.e., where training set samples belonging to more than one class are present. It may be convenient to distinguish between classifications which are unreliable because a sample is of type a) or b). To this end, let us define two reliability parameters, say ψ_a and ψ_b, whose values vary in the interval [0,1] and quantify the reliability of a classification from the two different points of view. Values near to 1 will characterize very reliable classifications, while low parameter values will be associated with classifications unreliable because the considered sample is of type a) or b). A parameter ψ providing a comprehensive measure of the reliability of a classification can result from the combination of the values of ψ_a and ψ_b. We have chosen the form $\psi = \min\{\psi_a, \psi_b\}$. This is certainly a conservative choice because it implies that, for a classification to be unreliable, just one reliability parameter needs to take a low value, regardless of the value assumed by the other one. By definition, the ideal reliability parameter should assume a value equal to 1 for all the correctly classified samples and a value equal to 0 for all the misclassified samples. However, this will almost never happen in real cases. The operative definition of ψ requires the classifier to provide an output consisting of a vector the values of whose elements make it possible to establish the class a sample belongs to.

As regards the MLP classifier, it can be shown [10] that an effective definition of the reliability parameter ψ_a can be $\psi_a = O_{win}$ where O_{win} is the output of the winner neuron, while a suitable reliability parameter for the case b) is $\psi_b = O_{win} - O_{2win}$. In conclusion, the classification reliability of the MLP classifier can be measured by:

$$\psi = \min\{\psi_a, \psi_b\} = \min\{O_{win}, O_{win} - O_{2win}\} = O_{win} - O_{2win} = \psi_b$$

For an LVQ classifier, the values of the elements of the output vector give the distances of an input sample X from each of the prototypes W_i, i=1,...,M, with M

generally greater than the number N of classes. Therefore, the winner neuron is the one having the minimum output value $O_{win} = \min_i \{O_i\} = \min_i \{d(W_i, X)\}$. With this assumption, a convenient form for the first parameter can be $\psi_a = \max\{1 - O_{win}/O_{max}, 0\}$ where O_{max} is the highest value of O_{win} among those relative to all the samples of the training set. For the case b), we have adopted the definition $\psi_b = 1 - O_{win}/O_{2win}$, where O_{2win} is the value of the output neuron having the second lowest distance from the input sample. The classification reliability for the LVQ classifier is thus given by:

$$\psi = \min\{\psi_a, \psi_b\} = \min\left\{\max\left\{1 - \frac{O_{win}}{O_{max}}, 0\right\}, 1 - \frac{O_{win}}{O_{2win}}\right\} \quad (1)$$

The same considerations hold for the NN classifier. In fact it assigns the input sample X to the class including the reference graph having the smallest distance from X. The only differences are that the value of O_{max} has to be computed on a set different from the reference set and O_{2win} is the distance between X and the reference graph having the second smallest distance from X, among all those belonging to a class different from that of O_{win}. Therefore, the classification reliability for the NN classifier is again given by equation (1).

4 Combining Criteria

Most of the combining criteria proposed in the framework of the multi-expert approach use the confidence degree assigned by an expert to each classification it performs. To evaluate the confidence degree of the vote given by the k-th expert, the most common choice [1] is the classification confusion matrix E^k whose generic element e_{ij}^k represents the number of times the k-th expert assigns to the j-th class a sample belonging to the i-th class, divided by the total number of samples belonging to the i-th class.

To investigate the influence of the set used to compute the confusion matrix on the obtained performance, the values of the elements of E^k were computed once using the training set and then using a set of data different from both the training set and the test set (see next Section).

The following combining criteria were used:

1) *Majority Voting* (*MV*): each expert votes for one class and the estimated (i.e., the actually assigned) class is the one voted by the majority. If more classes obtain the same number of votes, the values e_{ii}^k are used for tie breaking, i.e. the vote of each expert is weighted by the number representing the reliability of that expert when it assigns a sample to the class it is voting for.

2) *Bayesian Combination* (*BC*): the estimated class is the one which maximizes the a posteriori probability. The probability that a sample belongs to the i-th class when the k-th expert assigns it to the j-th class is assumed to be $e_{ij}^k \Big/ \sum_{i=1}^{n} e_{ij}^k$.

3) *Dempster-Shafer evidential reasoning (DS)*: this criterion is based on the Dempster-Shafer theory [11]. According to it, we define for each expert, the "belief" in every possible subset A of the set $\Theta = \{A_1, A_2, \ldots, A_m\}$, where A_i is a proposition representing the fact that a sample is assigned to the i-th class by the considered expert. The belief bel(.) is calculated from a function, called basic probability assignment, which is denoted $m(.)$, by using the equation

$$\text{bel}(A) = \sum_{B \subseteq A} m(B) \tag{2}$$

where B is any subset of A. Obviously, we have $\text{bel}(A_i) = m(A_i)$ and $\text{bel}(\Theta)=1$. In our case, when the k-th expert votes for the i-th class, we consider $m(A_i) = e_{ii}^k$ and $m(\Theta) = 1 - e_{ii}^k$. The values $m(A)$ supplied by each expert are combined via the Dempster rule, and the values $\text{bel}(A_i)$ are calculated using the equation (2). The estimated class is the one that maximizes the value of $\text{bel}(A_i)$.

The criteria based on the confidence degree were compared with a fourth criterion based on the reliability parameter:

4) *Majority Voting Using the Reliability Parameters (RV)*: this criterion differs from the first one only for the values used for tie breaking: in this case, in fact, the reliability parameters defined in the previous section are used.

5 Experimental Results and Conclusions

All the tests were performed using the NIST database 19 [3], which contains 8 sets of images extracted from 3699 Handwriting Sample Forms and digitized at 300 dpi. In particular, we used the sets hsf_3 and hsf_4. Only digits were considered.

The set hsf_3 was split in two sets: a training set (TRS), composed of 34,644 samples, used for training the MLP-NS, MLP-H and LVQ-NS experts, and a so called training-test set (TTS) made of 29,252 samples. As already mentioned, a subset of TRS (8000 samples) was assumed as reference set for the NN-S expert. TTS was used both to compute the confusion matrices and to establish the number of cycles for stopping the learning phase of the MLP-NS, MLP-H and LVQ-NS experts, in order to avoid the overtraining phenomenon [8].

The set hsf_4, made of 58,646 samples, was adopted as test set (TS).

The MLP-NS and MLP-H experts have been trained performing 5000 learning cycles, while for the LVQ-NS expert 2000 learning cycles were performed. The recognition rates obtained by each single expert on the considered sets are reported in Tab. 1.

Expert	TRS	TTS	TS
MLP-NS	97.96	96.52	88.53
LVQ-NS	98.80	96.66	85.90
MLP-H	95.56	94.59	85.63
NN-S	--	90.98	84.11

Tab. 1. Recognition rates obtained by the single experts on TRS, TTS and TS.

Eleven different multi-experts have been considered: 1 of them combines 4 experts, 4 of them combine 3 experts, and each of the remaining 6 combines 2 experts. The multi-experts have been designed so as to test all the significant combinations of experts, each obtained by pairing a classifier with a descriptor. The considered combinations of experts and the experimental results obtained with them are summarized in Tab. 2.

Let us remind that the recognition systems considered have been selected not because they have an outstanding performance on the used database, but in order to perform the test on systems adopting different description and classification paradigms. However the recognition rates obtained are neither low, considering the quality of the characters in the data base. The use of the reliability parameter allows to improve the recognition rates for all the considered multi-expert systems. The improvement is more significant when the number of experts is equal to two: in this case, in fact, the need of tie breaking is more frequent than in presence of three or more experts. The recognition improvement achieved by the multi-experts when using the combining rule RV, although limited to a few percent, should be considered relevant, since it depends only on the fact that the reliability parameter has been used while the combining rules and the experts have been fixed.

Multi-Expert				**Combining Rule**			
MLP-NS	**MLP-H**	**LVQ-NS**	**NN-S**	***RV***	***BC***	***DS***	***MV***
X	X	X	X	92.19	91.96	91.68	91.62
X	X	X		90.56	90.35	90.23	90.21
X	X		X	91.73	91.70	90.79	90.79
X		X	X	91.68	91.30	90.78	90.77
	X	X	X	91.30	91.21	90.50	90.50
X	X			89.49	87.91	87.79	87.78
X		X		89.49	87.19	87.23	87.23
X			X	90.74	89.41	88.11	88.11
	X	X		88.15	87.19	86.31	86.31
	X		X	88.24	87.83	86.53	86.52
		X	X	88.40	87.02	86.09	86.08

Tab. 2. Recognition rates obtained by each multi-expert as a function of the combining rule. Values in parentheses refer to the case in which the confusion matrices are computed on the training set instead of the training-test set.

This is still more true, if it is considered that the values shown in Tab. 2 are the result of the average over all the classes, while the improvement due to the reliability parameter is not uniformly distributed among classes. It has been verified that there is a smaller improvement for classes whose samples do not exhibit a large shape variability (e.g., the class of the 1s) while the improvement is significantly higher than the average for classes whose samples are quite different from each other, like the class of the 8s.

As regards the set of data used to evaluate the confusion matrix it has been noted that the use of a confusion matrix computed on TRS makes the recognition

performance worst for almost all the multi-expert systems and particularly for the combinations including an LVQ-NS expert. This is due to the fact that for this expert the difference between the recognition rates on TRS and TS is greater than for the other experts. The use of the reliability parameter, whose value does not depend on the choice of a specific set of data, allows to overcome the problem, existing when computing the confusion matrix, of selecting a set of data different from the training set, but adequately representative of the real world.

Finally it is worth noting that to determine the performance of a multi-expert system another important factor is the variety of the component experts. In fact, all the combination with the NN-S expert achieve the best performance; this confirms that the selection of experts as much as possible complementary as regards both description and classification methods can significantly improve the performance of a multi-expert system.

References

[1] L. Xu, A. Krzyzak, C.Y. Suen, "Method of Combining Multiple Classifiers and Their Application to Handwritten Numeral Recognition", *IEEE Trans. on Systems, Man and Cybernetics*, vol. 22, no. 3, pp. 418-435, January 1992.

[2] S.-B. Cho, J.H. Kim, "Combining Multiple Neural Networks by Fuzzy Integral for Robust Classification", *IEEE Trans. on Systems, Man and Cybernetics*, vol. 25, no. 2, pp. 380-384, February 1995.

[3] P.J. Grother, NIST Special Database 19, Technical Report, National Institute of Standards and Technology, 1995

[4] L.P. Cordella, C. De Stefano, M. Vento, "A Neural Network Classifier for OCR using Structural Descriptions", *Machine Vision and Applications*, no. 8, pp. 336-342, 1995.

[5] P. Foggia, C. Sansone, F. Tortorella, and M. Vento, "Character Recognition by Geometrical Moments on Structural Decompositions", *Proc. 4th Int. Conf. on Document Analysis and Recognition*, to appear.

[6] D.E. Rumelhart, J.L. Mc Clelland, *Parallel Distributed Processing Explorations in the Microstructure of Cognition*, Vol.1: Foundations. MIT Press, Cambridge, Mass, 1986.

[7] T. Kohonen, "The Self-Organizing Map," *Proc. of the IEEE*, vol. 78, no. 9, pp. 1464-1480, September 1990.

[8] R. Hecth-Nielsen, *Neurocomputing*, Addison-Wesley, Reading (MA), 1990.

[9] C. De Stefano, P. Foggia, F. Tortorella, M. Vento, "A Distance Measure for Structural Descriptions using Circle Arcs as Primitives" in *Proc. 13th Int. Conf. on Pattern Recogn.*, IEEE Comp. Soc. Press, Vol. II, pp. 290-294, 1996.

[10] L.P. Cordella, C. De Stefano, F. Tortorella, M. Vento, "A Method for Improving Classification Reliability of Multilayer Perceptrons", *IEEE Trans. on Neural Networks*, vol. 6, no. 5, pp. 1140-1147, September 1995.

[11] J. Gordon, E.H. Shortliffe, "The Dempster-Shafer Theory of Evidence", in B.G. Buchanan, E.H. Shortliffe (Eds.), *Rule-Based Expert Systems*, Addison-Wesley, pp. 272-292, 1984.