# Optimizing the Error/Reject Trade-off for a Multi-expert System Using the Bayesian Combining Rule

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, foggiapa, carlosan, ftortor, vento}@unina.it - WWW: http://amalfi.dis.unina.it

**Abstract.** Recently, in the framework of Pattern Recognition, methods for combining several experts (Multi-Expert Systems, MES) in order to improve the recognition performance, have been widely investigated. A main problem of MES is that the combining rule should be able to take the right classification decision even when the experts disagree. Anyway, in critical cases, a reject decision is convenient to reduce the risk of an error. Up to now, the problem of defining a reject rule for a MES has not been systematically explored.

We propose a method for determining the best trade-off between error rate and reject rate depending on the considered application domain, i.e. by taking into account the costs attributed, for the specific application, to misclassifications, rejects and correct classifications. Even though the method has general validity, in this paper its application to a MES using the Bayesian combining rule is presented.

## 1 Introduction

The idea of combining various experts for improving the classification rate of a recognition system, has been recently widely investigated. The rationale lies in the assumption that, by suitably combining the results of a set of experts according to a rule (combining rule), the weakness of each single expert can be compensated without losing the strength of each of them, and the obtained performance can result better than that of any single expert [1]. The successful implementation of a multiple-expert system (MES) implies the use of as much as possible complementary experts, and 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.

Preliminary experimental results encouraged this approach, and various research groups concentrated the attention on its different aspects [2,3,4].

The main problem of the approach is to find a combining rule able to solve the conflicts, i.e. to take the right classification decision even when the experts disagree; in this case, the final decision of the combiner cannot be always considered reliable. Very crucial situations for the combiner are the ones in which either a sample is attributed to two or more classes with a comparable likelihood, or no class can be considered sufficiently reliable. In these conditions it would be desirable to take a decision about the advantage of rejecting a sample (i.e., not assigning it to a class), instead of running the risk of misclassifying it. In practice, this advantage can only be evaluated by taking into account the requirements of the specific application domain. In fact, there are applications for which the cost of a misclassification is very high, so that a high reject rate is acceptable provided that the misclassification rate is kept as low as possible; a typical example could be the classification of medical images in the framework of a pre-screening for early cancer detection. In other applications it may be desirable to assign every sample to a class even at the risk of a higher misclassification rate. Let us consider, for instance, the case of applications in which the output of the system has to be submitted to an extensive manual post-processing. Between these extremes, a number of applications can be characterized by intermediate requirements. Thus a wise choice of the reject rule allows to tune the recognition system to the given application.

Up to now, the problem of defining a reject option for a MES has not been systematically explored. Simple heuristic reject rules are in fact used together with some of the combination rule presented in the literature. For example, one of the simplest combining rule, the "Majority Voting" [2], assigns the input sample the class for which a relative or absolute majority of experts agree, and rejects the sample in case two or more classes receive the same number of votes.

More sophisticated combining rules introduce a measure of the reliability associated to the response of each expert. In the "Weighted Voting" rule [2,3], the votes of all the experts are collected and the input sample is assigned to the class for which the sum of the votes, each weighted by the estimated reliability of the corresponding expert, is highest. Generally, weighted voting methods either do not introduce reject criteria, or, as suggested in [4], obtain the reject by fixing a threshold on the minimum value of the weighted vote and/or on the minimum tolerable difference between the highest vote and the second highest value. In this case, however, the threshold is not assigned by considering the requirements of the domain, and the behavior of the whole system could be not adequate for the considered application.

With reference to a combiner using the BKS rule [5], Suen and Huang propose a method for determining the reject threshold, once the desired error and reject rates have been assigned. Since this is made by using the BKS, this approach cannot be easily extended to other MES architectures.

In this paper we introduce a method for determining the best trade-off between error rate and reject rate in the considered application domain, for a MES using the Bayesian Combining rule [4], by taking into account the costs attributed, for the specific application, to misclassifications, rejects and correct classifications. Note that the last cost is actually a gain, but the term cost is used for the sake of notation uniformity.

The method is based on the estimate of the reliability of each classification act of the MES, and determines the optimal reject threshold, once the domain-dependent costs have been assigned. If the reliability is greater than the threshold, the decision of the MES is considered acceptable, otherwise the input sample is rejected.

A similar approach has been already followed for determining the reject threshold in the case of a single classifier [6]. It is worth pointing out that in our case, unlike other methods [7], no a priori knowledge about the probability distribution of the class populations is needed.

The definition of the parameter for measuring the classification reliability of the decision of the MES is made in the case of the Bayesian Combining Rule.

## 2 Combining Criteria

Many ways to combine classifier decisions, and thus to organize a MES, have been proposed in the recent past. Some of them are based on heuristic approaches, like voting or ranking, while others are founded on more formalized theoretical bases, like those based on the Dempster-Shafer evidence theory or on statistical methods [2]. Among these, the Bayesian Combining (*BC*) rule is very often used because it is based on a well settled mathematical framework and is simply applicable to several MES architectures. It can be applied even in case that each expert participating to the combination provides only the class assigned to the input sample. This is obviously the minimal information supplied by a classification system. From an operative point of view, the BC rule estimates the *a posteriori* probability that the input sample belongs to a generic class $C$ and selects the class with the highest post-probability. In the hypothesis of independence of the experts and that the *a priori* probability is the same for all the classes, the output $Y$ of the MES with the $BC$ rule is given by:

$$Y = \arg\max_{i} \prod_{k=1}^{M} p_k^i, \qquad 1 \le i \le n \tag{1}$$

where $p_k^i$ is the post-probability assigned to the class $C_i$ by the $k$-th expert, given the input sample $x$, $M$ is the number of the experts and $n$ the number of the classes.

There are several classifier paradigms which allow to obtain the post-probabilities $p_i^k$ (e.g., some classifiers based on a neural network [8]), so that the BC rule can be directly applied by using equation (1). This feature, however, is not shared by all the classifier architectures: in the most general case, the only information provided by an expert is the one specifying the most probable class an input samples belongs to. Therefore, in order to employ also in this case the BC rule, it is necessary to calculate the probability $P(x \in C_i \mid Y_k = o_k)$, i.e. the probability that the sample $x$ belongs to the class $C_i$ given that the output $Y_k$ of the $k$-th expert is equal to $o_k$ (the index specifying the class). In this way, if both the hypotheses of conditional independence among classifiers and of equiprobability of the classes hold, the output of the MES is given by:

$$Y = \arg\max_{i} \prod_{k=1}^{M} P(x \in C_i | Y_k = o_k), \quad 1 \le i \le n \tag{2}$$

An effective estimate for $P(x \in C_i \mid Y_k = o_k)$ can be obtained on the basis of the output of an expert, by taking into account its performance on a training set [4]. In particular, let us consider the classification confusion matrix $E^k$ for the $k$-th expert, whose generic element $e_{i,j}^k$ $(1 \le i,j \le n)$ represents the percentage of samples of the training set which belong to the $i$-th class and are assigned by the $k$-th expert to the $j$-th class. It is possible to show that:

$$P(x \in C_i \mid Y_k = o_k) \cong e^k_{i,o_k} \Big/ \sum_{h=1}^{n} e^k_{h,o_k} \tag{3}$$

By using the approximation in (3), we can rewrite equation (2), thus obtaining:

$$Y = \arg\max_i \prod_{k=1}^{M} \left( e^k_{i,o_k} \Big/ \sum_h e^k_{h,o_k} \right) \tag{4}$$

It is worth noting that in this way only the information specifying the class of belonging (*winning class*) is employed. Obviously, this is sufficient to establish the most likely class for an input sample, but a more careful look at the distribution of the values of the post-probabilities of the other classes could provide additional information about the classification reliability of the MES. This aspect is commonly disregarded in the definition of a combining rule, while it could significantly improve the performance of the whole classification system, especially in complex application domains. In the next sections we will show how to use such information for defining some reliability parameters and employing them in an optimal reject rule for a MES using the BC rule.

## 3 Classification Reliability and Reject Rule

Classification reliability can be expressed by associating a reliability parameter to every decision taken by the MES. Its quantitative evaluation requires the detection of situations which can give rise to unreliable classifications. The low reliability of a classification is generally due to one of the following situations: a) there is a diffused disagreement among the experts about the class to which the sample should be assigned and thus there is no class whose value of the post-probability is sufficient to judge the classification reliable; b) the experts part into groups each agreeing on a different class, but the values of the corresponding post-probabilities are so similar that there is not a clear overwhelming class.

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 parameters, say $\psi_a$ and $\psi_b$, whose values vary in the range [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). For the operative definitions of $\psi_a$ and $\psi_b$ (which will be referred to as reliability parameters), let us denote with $\pi_1$ the value of the post-probability associated to the winning class and with $\pi_2$ the value of the second maximal post-probability. A suitable definition for the reliability parameters is:

$$\psi_a = \pi_1 \qquad \text{and} \qquad \psi_b = 1 - (\pi_2/\pi_1) \tag{5}$$

In this way, if the value of $\psi_a$ is low, the corresponding classification is characterized by a weak post-probability and thus should be regarded as unreliable. Similar considerations holds for low $\psi_b$ value: in this case, however, there are more classes resulting equally probable and thus a reliable decision cannot be taken.

A parameter $\psi$ providing a comprehensive measure of the reliability of a classification can result from the combination of the values of $\psi_a$ and $\psi_b$:

$$\psi = \psi(\psi_a, \psi_b) \tag{6}$$

In this way, it is possible to judge about the reliability of the decision of the MES on the basis of a single value. There are several ways to combine the reliability parameters. A review of combination operators can be found in [9]. In the following some of the operators proposed in [9] will be used.

In classification problems regarding real applications, finding a reject rule which achieves the best trade-off between error rate and reject rate is undoubtedly of practical interest. The reject rule we propose for a BC-based MES compares the classification reliability $\psi$ with a suitably determined threshold $\sigma$: the classification is considered acceptable if the reliability value is greater than the threshold, otherwise the input sample is rejected. The reject rule is optimal with reference to the environment in which the MES works: in fact, the threshold $\sigma$ is computed by maximizing a function $\mathcal{P}$ which measures the MES classification effectiveness in the considered application domain. This requires to quantitatively estimate the consequences of the classification result in the particular domain: the cost of a misclassification is generally attributed by considering the burden of locating and possibly correcting the error or, if this is impossible, by evaluating the consequent damage. The cost of a reject is that of a new classification using a different technique.

To operatively define the function $\mathcal{P}$, let us call $R_c$ the percentage of correctly classified samples (also referred to as recognition rate), $R_e$ the misclassification rate (also called error rate), and $R_r$ the reject rate; moreover, let $R_c^0$ and $R_e^0$ indicate respectively the recognition rate and the error rate when the classifier is used at 0-reject. If we assume for $\mathcal{P}$ a linear dependence on $R_c$, $R_e$ and $R_r$, its expression is given by:

$$\mathcal{P}(R_c, R_e, R_r) = C_c(R_c - R_c^0) - C_e(R_e - R_e^0) - C_r R_r \tag{7}$$

In other words, $\mathcal{P}$ measures the actual effectiveness improvement when the reject option is introduced, independently of the absolute performance of the MES at 0-reject. The three quantities $C_e$, $C_r$ and $C_c$ respectively denote the cost of each error, the cost of each rejection and the gain of each correct classification, for a given application. The linear dependence assumption has been made mainly to simplify the illustration of the method and does not affect its generality; in [6] it is shown how the method can be extended to the case of a function $\mathcal{P}$ of generic form.

Since $R_c$, $R_e$ and $R_r$ depend on the value of the reject threshold $\sigma$, $\mathcal{P}$ is also a function of $\sigma$. To highlight such dependence, let $D_c(\psi)$ and $D_e(\psi)$ be, respectively, the occurrence density curves of correctly classified and misclassified samples as a function of the value of $\psi$. By definition, the integrals of $D_c(\psi)$ and $D_e(\psi)$, extended to the interval $[\psi_1, \psi_2]$ respectively provide the percentage of correctly classified and misclassified samples having values of $\psi$ ranging from $\psi_1$ to $\psi_2$. Their trend should be such that the majority of correctly classified samples is found for high values of $\psi$, while misclassified samples are more frequent for low values of $\psi$ (see Fig. 1).

The percentages of correctly classified and misclassified samples which are rejected after the introduction of a reject threshold $\sigma$ are given by the gray areas. $R_c$ ($R_e$) represents the percentage of samples which are correctly classified (misclassified) after the introduction of the reject option.

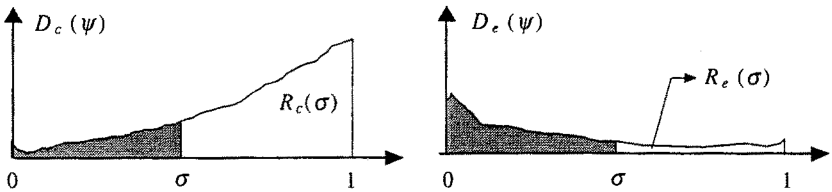


**Fig. 1:** Qualitative trends of the curves $D_c(\psi)$ and $D_e(\psi)$.

It is thus possible to directly evaluate the classification rate, the error rate and the reject rate for a given threshold $\sigma$:

$$R_c(\sigma)=\int_{\sigma}^{1} D_c(\psi)\,d\psi\;,\; R_e(\sigma)=\int_{\sigma}^{1} D_e(\psi)\,d\psi\;,\; R_r(\sigma)=\int_{0}^{\sigma}\left[D_c(\psi)+D_e(\psi)\right]d\psi \tag{8}$$

In this way $\mathcal{P}$ can be expressed as a function of $\sigma$:

$$\mathcal{P}(\sigma)=(C_e-C_r)\int_{0}^{\sigma} D_e(\psi)\,d\psi-(C_c+C_r)\int_{0}^{\sigma} D_c(\psi)\,d\psi \tag{9}$$

The optimal value $\sigma^*$ of the reject threshold $\sigma$ is the one for which the function $\mathcal{P}$ gets its maximum value.

In practice, the functions $D_c(\psi)$ and $D_e(\psi)$ are not available in their analytical form and therefore, for evaluating $\sigma^*$, they should be experimentally determined in tabular form on a set of labeled samples, adequately representative of the target domain. The optimal threshold $\sigma^*$ can be eventually determined by means of an exhaustive search among the tabulated values of $\mathcal{P}(\sigma)$. It can be simply shown that the location of the maximum for $\mathcal{P}(\sigma)$ depends on the ratio $C_n=(C_e-C_r)/(C_c+C_r)$, which will be referred to as *normalized cost.*

In order to correctly evaluate the improvement attainable with the reject option, it is worth introducing a parameter $P_n$ measuring the MES classification effectiveness normalized with respect to the maximum theoretical improvement (the *ideal* value of $\mathcal{P}$) $\mathcal{P}_{id}=(C_e-C_r)R_e^0$, which would be reached if all the errors were turned into rejects, without rejecting correctly classified samples. A suitable definition of $P_n$ is: $P_n=100\cdot\mathcal{P}(\sigma^*)/\mathcal{P}_{id}$. In this way, the trend of $P_n$ as a function of $C_n$ can give useful information about the improvement obtained for the MES as the application requirements vary.

## 4 Experimental Results and Discussion

The proposed method has been tested on the recognition of handprinted characters. Such application represents a critical recognition problem, since it is characterized by a high variability among the samples belonging to a same class and by partial overlaps among different classes.

In order to build up the MES to be employed, a set of experts has been considered. Each expert consists of a different pair (descriptor, classifier) so as to increase the probability that the experts are not correlated among each other. It is worth noting that the emphasis here is not much placed on the absolute performance of the description and classification techniques used, but on the improvement of the classification effectiveness achievable by introducing a reject option according to our method.

As regards the character description, we have considered four different methods. According to the first two, the character is described by means of a feature vector whose values are measures directly performed on the bit map. In particular, in the first case (*BM* description), the character is described through an 8 by 8 matrix of real numbers falling in the range [0,1]. It is obtained by superimposing an $8 \times 8$ grid on the character bitmap and by computing the average value of the pixels falling in each area. The matrix is finally coded as a 64 element vector. In the second case (*HA* description) we consider the Haar transform [10] of the character bitmap and build up a feature vector containing its first 64 coefficients. The other two methods are based on a structural representation of the character. This is obtained by means of a process which, starting from the bitmap, leads to a representation of the character in terms of circular arcs [11]. On this basis, we have defined a pure structural description (*ARG* description) and a hybrid description (*MA* description): according to the former, the character is described by means of an Attributed Relational Graph (ARG) whose nodes represent the component arcs (span, relative size and orientation) and whose branches represent the topologic relations between arcs. For the *MA* description, instead, geometric moments up to the 7th order are computed on the circular arcs constituting the structural representation of the character [12].

As regards the classifiers, we have employed a statistical classifier of the Nearest Neighbor type [13] and two neural networks. The first neural classifier was a Multi-Layer Perceptron (*MLP*) [14] with a single hidden layer of 30 neurons, while the second neural architecture was a Learning Vector Quantization (*LVQ*) [15] with a number of Kohonen neurons fixed to 7 for every classes.

We have considered five different experts: three of them employ the MLP classifier with the BM, HA and MA descriptions (let us denote these experts with MLP-BM, MLP-HA and MLP-MA); the fourth is constituted by the LVQ classifier with the BM description (LVQ-BM), and the last expert is a Nearest Neighbor classifier with the ARG description (NN-ARG). In this last case, we have used a metric defined in the ARG space [17].

All the tests were performed on the NIST database 19 [18]. For the tests, only digits were considered. As suggested by NIST, we used the set hsf_3 for training and the hsf_4 for testing. In particular, the set hsf_3 was split in two sets: a training set (TRS), composed of 34,644 samples, used for training the MLP-BM, MLP-HA, MLP-MA and LVQ-BM experts, and a so called training-test set (TTS) made of 29,252 samples. A subset of TRS (8000 samples) was assumed as reference set for the NN-ARG expert. TTS was used both to compute the confusion matrices and to establish the number of cycles for stopping the learning phase of the experts based on neural classifiers, in order to avoid the overtraining phenomenon [16]. The set hsf_4, adopted as test set (TS), is made of 58,646 samples; on this set the recognition rate of the MES made up of all the five experts was 93.82 at 0-reject. The performances of the single experts are reported in table 1.

| Expert | TRS | TTS | TS |
|---|---|---|---|
| MLP-HA | 99.47 | 97.77 | 90.47 |
| MLP-BM | 97.96 | 96.52 | 88.53 |
| MLP-MA | 95.56 | 94.59 | 85.63 |
| LVQ-BM | 98.80 | 96.66 | 85.90 |
| NN-ARG | -- | 90.98 | 84.11 |

**Table 1:** Recognition rates obtained by each single expert on TRS, TTS and TS.

As regards the cost coefficients, we assumed $C_c = 1$, while for $C_r$ and $C_e$ the pairs (5,9), (4,12), (4,15), (3,15), (3,18) were selected, with a normalized cost $C_n$ ranging from 0.67 to 3.75.

To combine the reliability parameters, four operators were considered with different peculiarities (see [9] for more details): $\psi_{min} = \min\{\psi_a, \psi_b\}$, $\psi_{med} = \frac{\psi_a + \psi_b}{2}$, $\psi_{max} = \max\{\psi_a, \psi_b\}$, $\psi_{sym} = \frac{\psi_a \psi_b}{1 - \psi_a - \psi_b + 2\psi_a \psi_b}$.

The following tables present the results obtained after the introduction of the reject option for different values of $C_n$ and for each of the considered operators. In particular, table 2 presents the values of the optimal threshold $\sigma^*$, while in table 3 $R_c$, $R_e$ and $R_r$ are reported.

| $C_n$ | $\psi_{min}$ | $\psi_{med}$ | $\psi_{max}$ | $\psi_{sym}$ |
|---|---|---|---|---|
| 0.67 | 0.000 | 0.000 | 0.000 | 0.000 |
| 1.60 | 0.000 | 0.000 | 0.000 | 0.000 |
| 2.20 | 0.000 | 0.000 | 0.991 | 0.999 |
| 3.00 | 0.000 | 0.000 | 0.991 | 0.999 |
| 3.75 | 0.999 | 0.999 | 0.997 | 0.999 |

**Table 2:** The values of $\sigma^*$ as a function of $C_n$, using the different reliability parameters.

As it is expected, when the value of $C_n$ increases, the recognition rate slightly decreases; this effect is balanced by a decrement of the error rate that leads to an overall improvement of the effectiveness of the MES. To properly characterize the MES effectiveness achievable with the introduction of the reject option, we have reported in fig. 2 the trend of $P_n$ with respect to $C_n$. It is worth noting that the effectiveness of the MES has an increasing trend with respect to $C_n$ and reaches its maximum increase (over 23%) in correspondence of the use of the $\psi_{max}$ operator.

The advantage attainable by exploiting the reject option can be made still more evident by considering the relative variation of classification and misclassification rates, with respect to the 0-reject case, as a function of $C_n$ (Table 4). It can be seen that, for high values of $C_n$, from 34% to 58% of the samples previously misclassified are now rejected, while the corresponding amount of correctly classified samples which are now rejected ranges from 5% to 9%.

| | $\psi_{min}$ | | | $\psi_{med}$ | | | $\psi_{max}$ | | | $\psi_{sym}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $C_n$ | $R_c$ | $R_r$ | $R_e$ | $R_c$ | $R_r$ | $R_e$ | $R_c$ | $R_r$ | $R_e$ | $R_c$ | $R_r$ | $R_e$ |
| 0.67 | 93.82 | 0.00 | 6.18 | 93.82 | 0.00 | 6.18 | 93.82 | 0.00 | 6.18 | 93.82 | 0.00 | 6.18 |
| 1.60 | 93.82 | 0.00 | 6.18 | 93.82 | 0.00 | 6.18 | 93.82 | 0.00 | 6.18 | 93.82 | 0.00 | 6.18 |
| 2.20 | 93.82 | 0.00 | 6.18 | 93.82 | 0.00 | 6.18 | 88.62 | 7.82 | 3.55 | 89.33 | 6.63 | 4.04 |
| 3.00 | 93.82 | 0.00 | 6.18 | 93.82 | 0.00 | 6.18 | 88.62 | 7.82 | 3.55 | 89.33 | 6.63 | 4.04 |
| 3.75 | 85.46 | 11.92 | 2.61 | 85.48 | 11.89 | 2.63 | 88.04 | 8.76 | 3.20 | 89.33 | 6.63 | 4.04 |

**Table 3**: Recognition, reject and error rates as a function of $C_n$ and different reliability parameters.

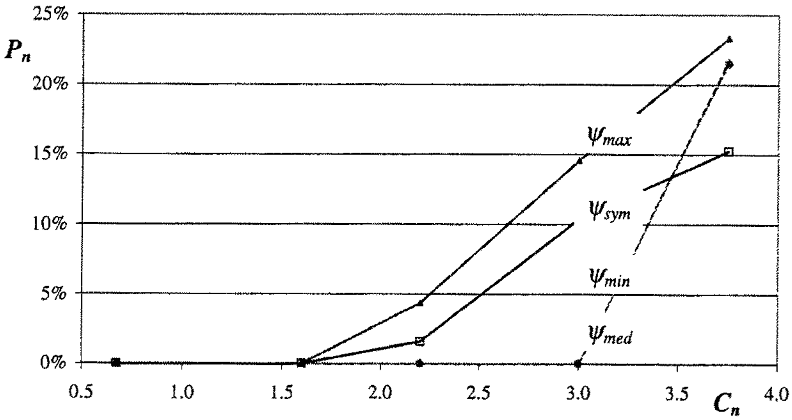


**Fig. 2**: The normalized effectiveness $P_n$ of the MES as a function of $C_n$.

In conclusion, it is important to recall that the overall improvement of the MES effectiveness is closely linked to the shape of the distributions $D_c$ and $D_e$, which, in turn, depend not only on the data but also on the definition of $\psi$. However, in real situations such as the one considered here, $D_c$ and $D_e$ are far from the ideal case since they overlap extensively. This makes them not separable and thus the attainable improvement of the MES effectiveness, whatever the definition of $\psi$, is lower than in the ideal case.

| | $\psi_{min}$ | | $\psi_{med}$ | | $\psi_{max}$ | | $\psi_{sym}$ | |
|---|---|---|---|---|---|---|---|---|
| $C_n$ | $\Delta R_c$ | $\Delta R_e$ | $\Delta R_c$ | $\Delta R_e$ | $\Delta R_c$ | $\Delta R_e$ | $\Delta R_c$ | $\Delta R_e$ |
| 0.67 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 1.60 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 2.20 | 0.00 | 0.00 | 0.00 | 0.00 | 5.54 | 42.56 | 4.79 | 34.63 |
| 3.00 | 0.00 | 0.00 | 0.00 | 0.00 | 5.54 | 42.56 | 4.79 | 34.63 |
| 3.75 | 8.91 | 57.77 | 8.89 | 57.44 | 6.16 | 48.22 | 4.79 | 34.63 |

**Table 4:** The relative variations of $R_c$ and $R_e$ as a function of $C_n$ and different reliability parameters.

## References

[1] C.Y. Suen, C. Nadal, R. Legault, T.A. Mai, and L. Lam, "Computer Recognition of Unconstrained Handwritten Numeral", *Proceedings of the IEEE*, vol. 80 no. 7, pp. 1162-1180, 1992.

[2] B. Ackermann, H. Bunke, "Combination of Classifiers on the Decision Level for Face Recognition", Technical Report IAM-96-002, Institut für Informatik und angewandte Mathematik, Universität Bern, January 1996.

[3] T.K. Ho, J.J. Hull, S.N. Srihari, "Decision Combination in Multiple Classifier Systems", *IEEE Transactions on PAMI*, vol. 16, n. 1, pp. 66-75, 1994.

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

[5] Y.S. Huang, C.Y. Suen, "A Method of Combining Multiple Experts for the Recognition of Unconstrained Handwritten Numerals", *IEEE Trans. on PAMI*, vol. 17, no. 1, pp. 90-94, 1995.

[6] 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, 1995.

[7] C. K. Chow, "On Optimum Recognition Error and Reject Tradeoff", *IEEE Trans. on Information Theory*, vol. 16, no. 1, January 1970.

[8] M.D. Richard, R.P. Lippman, "Neural Network Classifiers Estimates Bayesian *a posteriori* Probabilities", *Neural Computation*, vol. 3, no. 4, pp. 461-483, 1991.

[9] I. Bloch, "Information Combination Operators for Data Fusion: A Comparative Review with Classification", *IEEE Trans. on SMC - Part A*, vol. 26, no. 1, pp. 52-76, 1996.

[10] W.K. Pratt, *Digital Image Processing*, Wiley-Interscience, New York, 1991.

[11] 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.

[12] P. Foggia, C. Sansone, F. Tortorella, and M. Vento, "Character Recognition by Geometrical Moments on Structural Decompositions", in *Proc. of the 4$^{th}$ ICDAR*, Ulm (D), IEEE Computer Society Press, pp. 6-10, 1997.

[13] T.M. Cover, P.E. Hart, "Nearest Neighbor Pattern Classification", *IEEE Transactions on Information Theory*, vol. 13, pp. 21-27, 1967.

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

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

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

[17] C. De Stefano, P. Foggia, F. Tortorella, M. Vento, "A Distance Measure for Structural Descriptions Using Circle Arcs as Primitives" in *Proc. of the 13$^{th}$ ICPR*, Vienna, IEEE Computer Society Press, Vol. II, pp. 290-294, 1996.

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