EXPERIMENTS ON SOLVING MULTICLASS RECOGNITION TASKS IN THE BIOLOGICAL AND MEDICAL DOMAINS

Paolo Soda

Facoltà di Ingegneria, Università Campus Bio-Medico di Roma, Via Alvaro del Portillo 21, Roma, Italy

Keywords: Statistical Pattern Recognition, Decomposition Methods, One-per-class, Reliability Estimation, Classifier Ensembles.

Abstract: Multiclass learning problems can be cast as the task of assigning instances to a finite set of classes. Although in the wide variety of learning tools there exist some algorithms capable of handling polychotomies, many of the tools were designed by nature for dichotomies. In the literature, many techniques that decompose a polychotomy into a series of dichotomies have been proposed. One of the possible approaches, known as *one-per-class*, is based on a pool of binary modules, where each one distinguishes the elements of one class from those of the others. In this framework, we propose a novel reconstruction criterion, i.e. a rule that sets the final decision on the basis of the single binary classifications. It looks at the quality of the current input and, more specifically, it is a function of the reliability of each classification act provided by the binary modules. The approach has been tested on four biological and medical datasets and the achieved performance has been compared with the one previously reported in the literature, showing that the method improves the accuracies so far.

1 INTRODUCTION

Many supervised pattern recognition tasks can be cast as the problem of assigning elements to a finite set of classes or categories. Such tasks are referred to as binary learning, or dichotomies, when they aim at distinguishing instances of two classes, whereas they are named multiclass learning, or polychotomies, if there are more categories.

There is a huge number of applications that require multiclass categorization. Some examples are text classification, object recognition and support to medical diagnosis, to name a few.

In the literature numerous learning algorithms have been devised for multiclass problems, such as neural networks or decision trees. However it exists a different approach that is based on the reduction of the multiclass task into multiple binary problems, referred to as *decomposition method*. The problem complexity is therefore reduced trough the decomposition of the polychotomy in less complex subtasks. The basic observation that supports such an approach is that in the literature most of the available algorithms, which handle classification problems, are best suited to learning binary function (Dietterich and Bakiri, 1995; Mayoraz and Moreira, 1997). Different dichotomizers, i.e. the discriminating functions that subdivide the input patterns in two separated classes, perform the corresponding recognition task. To provide the final classification, their outputs are combined according to a given rule, usually referred to as *decision* or *reconstruction rule*.

In the framework of decomposition methods for classification, the various methods proposed to-date can be traced back to the following three categories (Dietterich and Bakiri, 1995; Mayoraz and Moreira, 1997; Jelonek and Stefanowski, 1998; Masulli and Valentini, 2000; Allwein et al., 2001; Crammer and Singer, 2002; Hastie and Tibshirani, 1998; Kuncheva, 2005).

The first one, called *one-per-class*, is based on a pool of binary learning functions, where each one separates a single class from all the others. The assignment of a new input to a certain class can be performed, for example, looking at the function that returns the highest activation (Dietterich and Bakiri, 1995; Masulli and Valentini, 2000).

The second approach, commonly referred to as *distribuited output code*, assigns a unique codeword, i.e. a binary string, to each class. If we assume that the string has n bit, the recognition system is composed by n binary classification functions. Given an unknown pattern, the classifiers provide a n-bit string that is compared with the codeword to set the

Soda P. (2008).
EXPERIMENTS ON SOLVING MULTICLASS RECOGNITION TASKS IN THE BIOLOGICAL AND MEDICAL DOMAINS.
In *Proceedings of the First International Conference on Bio-inspired Systems and Signal Processing*, pages 64-71
DOI: 10.5220/0001066500640071

Copyright © SciTePress

final decision. For example, the input sample is assigned to the class with the closest codeword, according to a distance measure, such as the Hamming one. In this framework, in (Dietterich and Bakiri, 1995) the authors proposed an approach, known as *error-correcting techniques* (ECOC), where they employed error-correcting codes as a distributed output representation. Their strategy was a decomposition method based on the coding theory that allowed obtaining a recognition system less sensitive to noise via the implementation of an error-recovering capability. Although the traditional measure of diversity between the codewords and the outputs of dichotomizers is the Hamming distance, other works proposed different measures. For example, Kuncheva in (Kuncheva, 2005) presented a measure that accounted for the overall diversity in the ensemble of binary classifiers.

The last approach is called n^2 classifier. In this case the recognition system is composed of $(n^2-n)/2$ base dichotomizers, where each one is specialized in discriminating respective pair of decision classes. Then, their predictions are aggregated to a final decision using a voting criterion. For example, in (Jelonek and Stefanowski, 1998) the authors proposed a voting scheme adjusted by the credibilities of the base classifiers, which were calculated during the learning phase of the classification.

This short description of the methods so far shows that the recognition systems based on decomposition methods are constituted by an ensemble of binary discriminating functions. On this motivation, for brevity such systems are referred to as Multy Dichotomies System (MDS) in the following.

In the framework of the one-per-class approach, we present here a novel reconstruction rule that relies upon the quality of the input pattern and looks at the reliability of each classification act provided by the binary modules. Furthermore, the classification scheme that we propose allows employing either a single expert or an ensemble of classifiers internal to each module that solves a dichotomy. Finally, the effectiveness of the recognition system has been evaluated on four different datasets that belongs to biological and medical applications.

The rest of the paper is organized as follows: in the next section we introduce some notations and we present general considerations related to the system configuration. Section 3 details the reconstruction method and section 4 describes and discusses the experiments performed on four different medical datasets. Finally section 5 offers a conclusion.

2 PROBLEM DEFINITION

2.1 Background

Let us consider a classification task on c data classes, represented by the set of labels $\Omega = \{\omega_1, \cdots, \omega_c\}$, with $c > 2$. With reference to the one-per-class approach, the multiclass problem is reduced into c binary problems, each one addressed by one module of the pool $M = \{M_1, \cdots, M_c\}$. We say that the module, or the dichotomizer, M_j is specialized in the jth class when it aims at recognizing if the sample x belongs either to the jth class ω_j or, alternatively, to any other class ω_i, with $i \neq j$. Therefore each module assigns to the input pattern $x \in \Re^n$ a binary label:

$$M_j(x) = \begin{cases} 1 & \text{if } x \in \omega_j \\ 0 & \text{if } x \in \omega_i,\ i \neq j \end{cases} \tag{1}$$

where $M_j(x)$ indicates the output of the jth module on the pattern x. On this basis, the codeword associated to the class ω_j has a bit equal to 1 at the jth position, and 0 elsewhere.

Notice that we have just mentioned *module* and not *classifier* to emphasize that each dichotomy can be solved not only by a single expert, but also by an ensemble of classifiers. However, to our knowledge, the system dichotomizers typically adopt the former approach, i.e. they are composed by one classifier per specialized module. For example, for their experimental assessments the authors used a a decision tree and a multi layer perceptrons with one hidden layer both in (Mayoraz and Moreira, 1997) and (Masulli and Valentini, 2000), respectively. The same functions were employed by Dietterich and Bakiri for the evaluation of their proposal in (Dietterich and Bakiri, 1995), whereas Allwein et al. used a Support Vector Machine (Allwein et al., 2001). A viable alternative to using a single expert is the combination of classifiers outputs solving the same recognition task. The idea is that the classification performance attainable by their combination should be improved by taking advantage of the strength of the single classifiers. Classifier selection and fusion are the two main combination strategies reported in the literature. The former presumes that each classifier has expertise in some local area of the feature space (Woods et al., 1997; Kuncheva, 2002; Xu et al., 1992). For example, when an unknown pattern is submitted for classification, the more accurate classifier in the vicinity of the input is selected to label it (Woods et al., 1997). The latter algorithms assume that the classifiers are applied in parallel and their outputs are combined to attain somehow a group of "consensus" (De Stefano et al., 2000; Kuncheva et al., 2001; Kittler et al., 1998). Typical fusion techniques include weighted mean, voting, correlation, probability, etc..

It is worth noticing that the modules, besides labelling each pattern, may supply other information typically related to the degree that the sample belongs to that class. In this respect, the various classification algorithms are divided into three categories, on the basis of the output information that they are able to provide (Xu et al., 1992). The classifiers of *type 1* supply only the label of the presumed class and, therefore, they are also known as experts that work at the *abstract* level. *Type 2* classifiers work at the *rank* level, i.e. they rank all classes in a queue where the class at the top is the first choice. Learning functions of *type 3* operate at the *measurement* level, i.e. they attribute each class a value that measure the degree that the input sample belongs to that class. If a crisp label of the input pattern is needed, we can use the maximum membership rule that assigns x to the class for which the degree of support is maximum (ties are resolved arbitrarily). Although abstract classifiers provide a n-bit string that can be compared with the codewords, decision schemes that exploit information derived from the classifiers working at the measurement level permit us to define reconstruction rules that are potentially more effective. Furthermore, if the module is constituted by a multi-experts system, the information supplied by the single classifiers can be used to compute a measure similar to that provided by measurement classifiers.

Since measurement classifiers can provide more information with respect to the other two types, we assume that only measurement experts constitutes our MDS. Therefore, the research focus becomes: "Given the individual decision $M_1(x), \cdots, M_c(x)$ and the degrees of membership of x to the different classes, how can we use such an information to set the final label?".

2.2 The Reconstruction Method

The reconstruction method addresses the issues of determining the final label of the input pattern x on the basis of the modules' decisions and, eventually, of information directly derived from their outputs. To present our method, let us introduce two auxiliaries quantities. The first, named *binary profile*, represents the state of the module outputs. It is a c-bit vector defined by:

$$\mathbf{M}(x) = [M_1(x), \cdots, M_j(x), \cdots, M_c(x)] \tag{2}$$

whose entries are the crisp labels provided by each module in the classification of sample x (see equation 1).

Since each block has a binary output, the 2^c possible bit combinations of the binary profile can be grouped into the following three categories:

(i) only one module classifies the sample in the class in which it is specialized;

(ii) more modules classify the sample in its own class;

(iii) none module classifies the sample in its own class.

In the first case, only one entry of $\mathbf{M}(x)$ is one; in the second more elements are one (at least two and no more than c), whereas in the last situation all the elements are zero. Such an observation naturally leads to distinguish these three cases using the summation over the binary profile. Indeed,

$$m = \sum_{j=1}^{c} M_j(x) = \begin{cases} 1, & \text{in case (i)} \\ [2,c], & \text{in case (ii)} \\ 0, & \text{in case (iii)} \end{cases} \tag{3}$$

where m therefore represents the number of modules whose outputs are 1.

The second quantity that we introduce is referred to as *reliability profile* and it is described by:

$$\psi(x) = [\psi_1(x), \cdots, \psi_j(x), \cdots, \psi_c(x)] \tag{4}$$

where each element $\psi_j(x)$ measure the reliability of the classification act on pattern x provided by the jth module. Note that the reliability varies in the interval $[0,1]$, and a value near 1 indicates a very reliable classification.

We deem that the estimation of the reliability of each classification act is a viable method to employ the information directly derived from the classifiers output since it has demonstrated its convenience, in other field also (De Stefano et al., 2000; Cordella et al., 1999).

Assuming that we determined both the binary and the reliability profiles, i.e. $\mathbf{M}(x)$ and $\psi(x)$ respectively, in the next section we will present the reconstruction rule.

3 RELIABILITY BASED RECONSTRUCTION

In this section we introduce the novel reconstruction strategy we propose in the paper. It chooses an output in any of the 2^c combinations of the binary profile. We deem that an accurate final decision can be taken if the reconstruction rule looks at the quality of the classification provided by the modules, i.e. at the reliability of their specific decisions. To our knowledge the application of such a parameter can not be found in the literature related to decomposition methods. Indeed, the papers of this field that used the information directly derived from the outputs of the base classifiers typically considered only the highest activation among the experts, e.g. the maximum output from a pool of neural networks. However, this measure cannot be regarded as a reliability parameter, since it has been demonstrated that it should be computed considering not only the winner output neurons but also the losers (Cordella et al., 1999).

Therefore, differently from the past, we propose a criterion that makes use of the reliability measure, i.e. of the reliability profile, named as *Reliability-based-Reconstruction (RbR)*. Denoting by s the index of the module that sets the final output $O(x) \in \Omega$, referred to as selected module for brevity in the following, the final decision is given by:

$$O(x) = \omega_s \tag{5}$$

with

$$s = \begin{cases} \arg\max_j (M_j(x) \cdot \psi_j(x)), \text{ if } m \in [1, c] \\ \arg\min_j (\overline{M_j(x)} \cdot \psi_j(x)), \text{ if } m = 0 \end{cases} \tag{6}$$

where $\overline{M_j(x)}$ indicates the negate output of the jth block.

The first row of this equation considers both cases (i) and (ii). Indeed, since in the first case all the modules agree in their decision, as a final output is chosen the class of the module whose output is 1. Conversely, in cases (ii) and (iii) the final decision is performed looking at the reliability of each modules' classifications. In case (ii), m modules vote for their own class, whereas the others $(c-m)$ indicate that x does not belong to their own class. To solve the dichotomy between the m conflicting modules we look at the reliability of their classification and choose the class associated to the more reliable one. In case (iii) $m = 0$, suggesting that all modules classify x as belonging to another class than the one they are specialized. In this case, the bigger is the reliability parameter $\psi_j(x)$, the less is the probability that x belongs to ω_j, and the bigger is the probability that it belongs to the other classes. These observations suggest finding out which module has the minimum reliability and then choosing the class associated to it as a final output.

Panel A of figure 1 shows the architecture of the proposed recognition system. The decision $M_j(x)$ and the reliability $\psi_j(x)$ supplied by each of the c modules are aggregated in the *reconstruction module* to provide the final decision $O(x)$. As observed in section 2.1, the use of an ensemble of classifiers in each module is a way to improve its discrimination capability. In this respect, the panel B of the same figure depicts a typical configuration of a multi-experts system. Notice that both the output of the kth classifier and its reliability, denoted as $V_k(x)$ and $\xi_k(x)$, respectively, can be given to the combination rule in order to label the input sample.

4 EXPERIMENTAL EVALUATION

In this section we first describe the datasets used to assess the performance of the reconstruction method and, second, we briefly discuss the configuration of the MDS modules. Third, we present a strategy to evaluate the classification reliability when the modules are constituted both by a single classifier and by an ensemble of experts, respectively. Finally, we report the experimental results.

4.1 Datasets

For our tests we use four datasets, described in the following and summerized in table 1.

Indirect Immunofluorescence Well Fluorescence Intensity. Connective tissue diseases are autoimmune disorders characterized by a chronic inflammatory process involving connective tissues. When they are suspected in a patient, the Indirect Immunofluorescence (IIF) test based on HEp-2 substrate is usually performed, since it is the recommended method. The interested reader may find a wide explanation of the IIF and its issues in (Kavanaugh et al., 2000; Rigon et al., 2007). The dataset consists of 14 features extracted from 600 patients sera collected at Università Campus Bio-Medico di Roma. The samples are distributed over three classes, namely positive (36.0%), negative (32.5%) and intermediate (31.5%). Previous results are reported in (Soda and Iannello, 2006) where the authors employed a multiclass approach, achieving an accuracy of 76% approximately.

Indirect Immunofluorescence HEp-2 cells staining pattern. This is a dataset with 573 instances represented by 159 statistical and spectral features. The samples are distributed in five classes that are representative of the main staining patterns exhibited by HEp-2 cells, namely homogeneous (23.9%), peripheral nuclear (21.8%), speckled (37.0%), nucleolar (8.2%) and artefact (9.1%). These patterns are related to one of the different autoantibodies that give rise to a connective tissue disease. For the details on these classes see (Rigon et al., 2007). On this dataset, we performed some tests adopting a multiclass approach, which exhibits a hit rate of 63.6% approximately, evaluated using the eightfold cross validation.

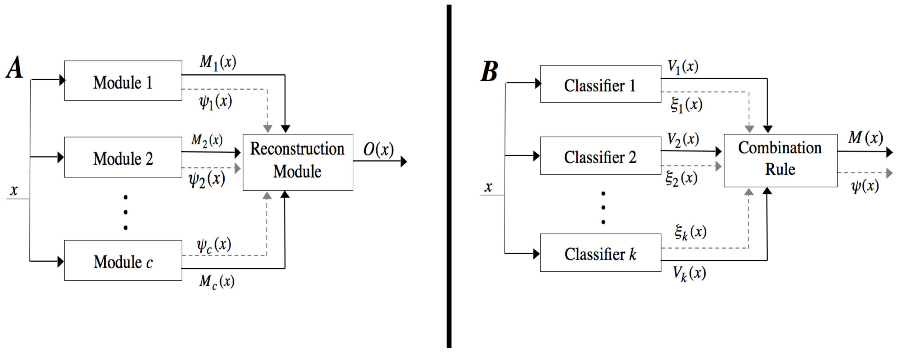

Figure 1: The system architecture, which is based on the aggregation of binary modules (panel ***A***), according to the one-per-class approach. Note that each module can be constituted by a multi-experts system, as depicted in the panel ***B***.

Lymphography. A database of lymph diseases was obtained from the University Medical Centre, Institute of Oncology, Ljubljana. It is composed by 148 instances described by 18 numeric attributes. There are four classes, namely normal (1.4%), metastases (54.7%), malign lymph (41.2%) and fibrosis (2.7%). The data are available within the UCI Machine Learning Repository[1] (Asuncion and Newman, 2007). Different approaches were used in the literature to address the recognition task. For instance, for Naive Bayes classifier and C4.5 decision tree the achieved performance was 79% and 77% respectively (Clark and Niblett, 1987), whereas induction techniques correctly classified the 83% of samples (Cheung, 2001).

Ecoli. The database is composed by 336 samples, described by a nine-dimensional vector and distributed in eight classes. Each class represents a localization site, which can be cytoplasm (42.5%), inner membrane without signal sequence (22.9%), periplasm (15.5%), inner membrane, uncleavable signal sequence (10.4%), outer membrane (6.0%), outer membrane lipoprotein (1.5%), inner membrane lipoprotein (0.6%) and inner membrane, cleavable signal sequence (0.6%). Again, the data are available within the UCI Machine Learning Repository (Asuncion and Newman, 2007). In (Jelonek and Stefanowski, 1998), the authors reported an accuracy that ranges from 79.7% up to 83.0%, achieved employing both a decision tree and a Multi Layer Perceptrons, respectively. In (Allwein et al., 2001), using many popular classification algorithms, such as the support-vector machines, AdaBoost, regression and decision-tree algorithms, the hit rate varies from 78.5% up to 86.1%.

4.2 MDS Configuration

The modules of the MDS are essentially composed by a single classifier or by an ensemble of classifiers. In both cases, as single expert we use k-Nearest Neighbour (kNN) or Multi-Layer Perceptron (MLP). For each dichotomy, we first select a subset of features that simplifies both the pattern representation and the classifier complexity as well as the risk of the incurring in the peaking phenomenon[2]. Then we carry out some preliminary tests to determine the best configuration of experts parameters, e.g. the number of neighbours for kNN classifier or the number of hidden layers, neurons per layer, etc., for the MLP network. Furthermore, when the module is constituted by an ensemble of experts we adopt a fusion technique to combine their outputs, namely the Weighted Voting (WV). In such a method the opinion of each expert about the class of the input pattern is weighted by the reliability of its classification. Since each expert deals with a binary learning task, to further present this scheme we can simplify the notation as follows. Denoting as $V_k(x)$ and as $\xi_k(x)$ the output and the classification reliability of kth classifier on sample x, the weighted sum of the votes for each of the two classes is given by:

$$W_h(x) = \sum_{k:V_k(x)=h} \xi_k(x), \text{ with } h = \{0,1\} \tag{7}$$

[1] For each dataset of this repository the users have access to a description of the application domain, to the features and to the ground truth.

[2] The peaking phenomenon is a paradoxical behaviour in which the added features may actually degrade the performance of a classifier if the number of training samples that are used to design the classifier is small relative to the number of features.

SCITEPRESS Science and Technology Publications

Table 1: Summary of the datasets used.

Database	Number of Samples	Number of Classes	Number of features	Avalaibility
IIF Well Fluorescence Intensity	600	3	14	Private
IIF HEp-2 cells staining pattern	573	5	159	Private
Lymphography	148	4	18	UCI
Ecoli	336	8	9	UCI

The output of the fusion of the jth module, $M_j(x)$, is defined by[3]:

$$M_j(x) = \begin{cases} 1 & \text{if } W_1(x) > W_0(x) \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Turning our attention to the configuration of the system in the experimental tests, notice that the modules that label the samples of the IIF Well Fluorescence Intensity and of lymphography datasets are composed by one classifiers. The modules that classify the samples of the HEp-2 cells and of the ecoli databases are constituted by kNN and MLP classifiers combined by the WV criterion.

4.3 Reliability Parameter

The approach described for deriving the final decision according to the RbR rule requires the introduction of a reliability parameter that evaluates the quality of the classification performed by each module, which can be composed by a single classifier or by an aggregation of classifiers (figure 1). In the former case its reliability ψ_j coincides with the one of the single classifier, i.e. ξ. In the latter case, each entry of the reliability profile generally depends on the combination rule adopted in the module, on the number k of composing experts and on their individual reliabilities ξ. Formally,

$$\psi_j(x) = \begin{cases} \xi(x), & \text{if } k = 1 \\ f(\xi_1(x), \cdots, \xi_i(x), \cdots, \xi_k(x)), & \text{if } k > 1 \end{cases} \tag{9}$$

where all the reliabilities are reported as function of the input pattern to emphasize that they are computed for each classification act.

In the rest of this section we first present two techniques to measure the reliability of kNN and MLP decisions, and then we introduce a novel method that estimates such parameter in the case of the application of the WV criterion.

A typical approach that measures the reliability of the decision taken by the single expert, i.e. ξ, makes use of the confusion matrix[4] estimated on the learning set. The drawback of this method is that all the patterns with the same label have equal reliability, regardless of the quality of the sample. Indeed, the average performance on the learning set, although significant, does not necessarily reflect the actual reliability of each classification act. To overcome such limitations we adopt an approach that relies upon the quality of the current input. To this end, we refer to the work presented in (Cordella et al., 1999), where the quality of the sample is related to its position in the feature space. In this respect, the low reliability of a recognition act can be traced back to one of the following situations: (a) in the feature space x is located in a region that is far from those associated with the various classes, i.e. the sample is significantly different from those present in the training set, (b) the point representing x lies in a region of the feature space where two or more classes overlap. These observations lead to introduce the parameters ξ_a and ξ_b that distinguish between the two situations of unreliable classification. Then, a comprehensive parameter ξ can be derived adopting the following conservative choice:

$$\xi = min(\xi_a, \xi_b) \tag{10}$$

Indeed, it implies that a low value for only one of the parameters is sufficient to consider unreliable the classification.

In the case of kNN classifiers, following (Cordella et al., 1999), the two parameters are defined are given by:

$$\xi_a = \max\left(1 - D_{min}/D_{max}, 0\right) \tag{11}$$

$$\xi_b = 1 - D_{min}/D_{min2} \tag{12}$$

where D_{min} is the smallest distance of x from a reference sample belonging to the same class of x, D_{max} is the highest among the values of D_{min} obtained for samples taken from the training-test set, i.e. a set that is disjoint from both the reference and the test set, D_{min2} is the distance between x and the reference sample with the second smallest distance from x among

[3] In case of tie, i.e. if $W_1(x)$ is equal to $W_0(x)$, the output $M_j(x)$ is set arbitrarily to zero. Note that it never occurred in all tests we performed.

[4] The confusion matrix reports for each entry (p,q) the percentage of samples of the class C_p assigned to the class C_q.

all the reference set samples belonging to a class that is different from that determining D_{min}.

In the case of MLP classifier, the two quantities are defined as follows:

$$\xi_a = N_{win} \tag{13}$$

$$\xi_b = N_{win} - N_{2win} \tag{14}$$

where N_{win} is the output of the winner neuron, N_{2win} is the output of the neuron with the highest value after the winner. From this definition, it is straightforward that $\xi = \xi_b$. For further details see (De Stefano et al., 2000).

When the jth module is composed by more than one classifier combined according to the WV rule, the reliability estimator considers again the situations which can give rise to an unreliable classification. In this respect, we need to introduce the following two auxiliary quantities:

$$\pi_1(x) = \max\left(\{\xi_k(x) | k : V_k(x) = M_j(x)\}\right) \tag{15}$$

$$\pi_2(x) = \max\left(\{\xi_k(x) | k : V_k(x) \neq M_j(x)\} \cup \{0\}\right) \tag{16}$$

where $\pi_1(x)$ and $\pi_2(x)$ represent the maximum reliabilities of experts voting for the winning class and for other classes (0 if all the experts agree on the winner class), respectively. Given these definitions, the reliability of the WV rule can be evaluated according to the following conservative choice:

$$\psi(x) = \min\left(\pi_1(x), \max\left(0, 1 - \pi_2(x)/\pi_1(x)\right)\right) \tag{17}$$

4.4 Results and Discussion

This section presents the experimental results that we achieved using the system described so far. To evaluate and then compare the results of this approach with those reported in the literature we perform eightfold and tenfold cross validation on the two IIF datasets, i.e. well fluorescence intensity and HEp-2 cells staining pattern, and on the other two databases, i.e. lymphography and ecoli, respectively.

The third column of table 2 shows the testing accuracies achieved on the four databases. To simply compare them with the past results, the second column of the same table summarizes the performance reported in literature. Turning our attention to the tests carried out on the first and on the second datasets, a relevant accuracy improvement can be observed. Indeed, the hit rate increases of 18.4% and of 12.3% in the case of the well fluorescence intensity and HEp-2 cells staining pattern databases, respectively. In our opinion, such an improvement is twofold motivated. On the one hand, the set of extracted features is more stable and more effective when we adopt a decomposition approach rather than a multiclass one. On the other hand, the reconstruction rule exhibits a very good capability of solving the disagreements between the specialized modules. Indeed, when the binary profile of the input sample $\mathbf{M}(x)$ differs from one of the possible codewords (i.e. $m = 0$ or $2 \leq m \leq c$), the decision is taken looking at the reliability profile $\psi(x)$, as presented in the formula 6. These considerations are strengthened by the observation of the performance attained in the classification of samples belonging to the two UCI datasets. Indeed, since they are benchmark datasets, any reported improvement is due to the recognition approach rather than to the use of a different features set. The tests on both the lymphography and ecoli datasets exhibit an accuracy better than the one reported to date. Indeed, for the former dataset the improvement ranges both from 6.9% up to 12.9% , whereas for the latter one it varies from 1.8% up to 9.4%. Therefore, also in these cases the MDS in combination with the RbR rule improves the recognition performance. Furthermore, it is worth noting that the approach seems independent of the modules' arrangement. The rationale lies in observing that in two of the four tests the MDS modules are constituted by a multi-experts system, whereas in the others they are composed by a single classifier (see the beginning of section 4). Consequently, the reliability ψ_j is measured according to a method that varies with the module configuration, as previously presented (see equations 10-17). Nevertheless, these variations do not affect the effectiveness of the recognition system. Therefore, we deem that the reconstruction rule is robust with respect to different reliability estimators.

5 CONCLUSIONS

In the framework of decomposition methods, we have presented a classification approach that reconstructs the final decision looking at the reliability of each classification act provided by all dichotomizers. Furthermore, the reconstruction rule does not depend on the configuration of each module, i.e. on its architecture. Such an observation is strengthened by the good performance achieved when both a single classifier and a fusion of experts constitute each module, respectively.

For all the four tested databases, the experimental results show that the proposed system outperforms the performance reported in the literature.

Future works are directed towards two issues. First, the test of the system on other public datasets and, second, the definition of reliability parameter of

Table 2: Testing accuracy achieved on the used datasets.

Database	Past Usage	MDS using RbR
IIF Well Fluorescence Intensity	75.9%	94.3%
IIF HEp-2 cells staining pattern	63.6%	75.9%
Lymphography	77% – 83.0%	89.9%
Ecoli	78.5% – 86.1%	87.9%

each decision taken by the MDS.

ACKNOWLEDGEMENTS

The author would like to thank the DAS s.r.l (`www.dasitaly.com`), which has funded this work.

REFERENCES

Allwein, E. L., Schapire, R. E., and Singer, Y. (2001). Reducing multiclass to binary: a unifying approach for margin classifiers. *J. Mach. Learn. Res.*, 1:113–141.

Asuncion, A. and Newman, D. J. (2007). UCI machine learning repository.

Cheung, N. (2001). Machine learning techniques for medical analysis. Master's thesis, University of Queensland.

Clark, P. and Niblett, T. (1987). Induction in noisy domains. In *Progress in Machine Learning–Proc. of EWSL 87*, pages 11–30.

Cordella, L., Foggia, P., and et. al. (1999). Reliability parameters to improve combination strategies in multi-expert systems. *Patt. An. & Appl.*, 2(3):205–214.

Crammer, K. and Singer, Y. (2002). On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.*, 2:265–292.

De Stefano, C., Sansone, C., and Vento, M. (2000). To reject or not to reject: that is the question: an answer in case of neural classifiers. *IEEE Trans. on Systems, Man, and Cybernetics–Part C*, 30(1):84–93.

Dietterich, T. G. and Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263.

Hastie, T. and Tibshirani, R. (1998). Classification by pairwise coupling. In *NIPS '97: Proc. of the 1997 Conf. on Advances in neural information processing systems*, pages 507–513. MIT Press.

Jelonek, J. and Stefanowski, J. (1998). Experiments on solving multiclass learning problems by n^2 classifier. In *10th European Conference on Machine Learning*, pages 172–177. Springer-Verlag Lecture Notes in Artificial Intelligence.

Kavanaugh, A., Tomar, R., and et al. (2000). Guidelines for clinical use of the antinuclear antibody test and tests for specific autoantibodies to nuclear antigens. *Am. Col. of Pathologists, Archives of Pathology and Lab. Medicine*, 124(1):71–81.

Kittler, J., Hatef, M., and et. al. (1998). On combining classifiers. *IEEE Trans. On Pattern Analysis and Machine Intelligence*, 20(3):226–239.

Kuncheva, L. I. (2002). Switching between selection and fusion in combining classifiers: an experiment. *IEEE Trans. on Systems, Man and Cybernetics, Part B*, 32(2):146–156.

Kuncheva, L. I. (2005). Using diversity measures for generating error-correcting output codes in classifier ensembles. *Patt. Recogn. Lett.*, 26(1):83–90.

Kuncheva, L. I., Bezdek, J. C., and Duin, R. (2001). Decision template for multiple classifier fusion: an experimental comparison. *Patt. Recognition*, 34:299–314.

Masulli, F. and Valentini, G. (2000). Comparing decomposition methods for classication. In *KES'2000, Fourth Int. Conf. on Knowledge-Based Intell. Eng. Systems & Allied Technologies*, pages 788–791.

Mayoraz, E. and Moreira, M. (1997). On the decomposition of polychotomies into dichotomies. In *ICML '97: Proc. of the 14th Int. Conf. on Machine Learning*, pages 219–226. Morgan Kaufmann Publishers Inc.

Rigon, A., Soda, P., Zennaro, D., Iannello, G., and Afeltra, A. (2007). Indirect immunofluorescence (IIF) in autoimmune diseases: Assessment of digital images for diagnostic purpose. *Cytometry - Accepted for Publication, February*.

Soda, P. and Iannello, G. (2006). A multi-expert system to classify fluorescent intensity in antinuclear autoantibodies testing. In *Computer Based Medical Systems*, pages 219–224. IEEE Computer Society.

Woods, K., Kegelmeyer, W., and Bowyer, K. (1997). Combination of multiple classifiers using local accuracy estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:405–410.

Xu, L., Krzyzak, A., and Suen, C. (1992). Method of combining multiple classifiers and their application to handwritten numeral recognition. *IEEE Trans. on Systems, Man and Cybernetics*, 22(3):418–435.