IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

Improved Pairwise Coupling Classification With Correcting Classifiers

Miguel Moreira † Eddy Mayoraz †

IDIAP–RR 97-09

October 1997

PUBLISHED IN
Tenth European Conference on Machine Learning (ECML'98)
Chemnitz, Germany, April 21-24 1998

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

† IDIAP—Dalle Molle Institute of Perceptual Artificial Intelligence, P.O. Box 592, CH-1920 Martigny, Switzerland, {miguel,mayoraz}@idiap.ch.

IDIAP Research Report 97-09

Improved Pairwise Coupling Classification With Correcting Classifiers

Miguel Moreira Eddy Mayoraz

October 1997

published in
Tenth European Conference on Machine Learning (ECML'98)
Chemnitz, Germany, April 21-24 1998

Abstract. The benefits obtained from the decomposition of a classification task involving several classes, into a set of smaller classification problems involving two classes only, usually called dichotomies, have been exposed in various occasions. Among the multiple ways of applying the referred decomposition, Pairwise Coupling is one of the best known. Its principle is to separate a pair of classes in each binary subproblem, ignoring the remaining ones, resulting in a decomposition scheme containing as much subproblems as the number of possible pairs of classes in the original task. Pairwise Coupling decomposition has so far been used in different applications. In this paper, various ways of recombining the outputs of all the classifiers solving the existing subproblems are explored, and an important handicap of its intrinsic nature is exposed, which consists in the use, for the classification, of impertinent information. A solution for this problem is suggested and it is shown how it can significantly improve the classification accuracy. In addition, a powerful decomposition scheme derived from the proposed correcting procedure is presented.

Keywords: Classification, decomposition into binary subproblems, pairwise coupling.

Acknowledgements: The authors are thankful to Prof. Alain Hertz for the valuable discussions that greatly contributed to the present work. The support of the Swiss National Science Foundation under grant FN 21-46974.96 is also gratefully acknowledged.

1 Introduction

The goal in automated learning consists in finding an approximation $\hat{F}$ of an unknown function F defined from an *input space* Ω onto an *output space* Σ, given a *training set* : $T = \{(\boldsymbol{x}^p, F(\boldsymbol{x}^p))\}_{p=1}^{P} \subset \Omega \times \Sigma$. When the output space is discrete and unordered, a *classification* problem is presented and the function $F : \Omega \rightarrow \{1, \ldots, K\}$ defines a K-partition of the input space into sets $F^{-1}(k)$ called *classes* and denoted ω_k.

The collection of learning algorithms available to solve classification problems originate in different domains such as: statistics (e.g. Bayesian classifiers, see [8]), logic (e.g.logical analysis of data [4, 1]), neural networks (e.g. perceptron algorithm [16], backpropagation [19]), artificial intelligence (e.g. decision trees [2, 14]). Among these, only those capable of handling multiclass problems are applied, in general, to solve problems where the number of classes exceeds two.

It is possible, however, to apply Boolean methods (i.e. those that can handle only two-class problems) to learn tasks where $K \gg 2$. In fact, different reasons motivate the decomposition of a large scale problem into smaller subproblems dealing with only two classes. On the one hand, some algorithms do not scale up nicely with the size of the training set. Others are not suited to handle a large number of classes. On the other hand, even when using an approach which can deal with large scale problems, an adequate decomposition of the classification problem into subproblems can be favorable to the overall computational complexity as well as to the generalization ability of the global classifier [17, 3, 20].

The use of a *decomposition scheme* allows the transformation of a K-partition $F : \Omega \rightarrow \{1, \ldots, K\}$, into a series of L bipartitions, $f_1, \ldots, f_L$. A *reconstruction method* is coupled with each decomposition scheme to make the fusion of the answers of all the L classifiers for a particular input in order to select one of the K classes. Among the simplest decomposition schemes frequently used, there is the *one-per-class* (OPC) and the *pairwise coupling* (PWC). A K-partition is decomposed by the former method into K bipartitions, each separating one class from all the others. The latter requires $\frac{1}{2}K(K-1)$ 2-class problems, one for each pair of classes, the bipartition for the pair (i, j) focusing on the separation of class ω_i from class ω_j and ignoring all other data. This paper concentrates on these two decomposition schemes, which are quite intuitive. A more sophisticated scheme is proposed in [5, 7, 10]; in ECOC, redundancy is explored in the decomposition as a way of increasing the error-correcting capability of the reconstruction. This method has served as base for other developments in the same framework. The authors in [11] present a similar scheme where the error-correcting component is kept, but the decomposition is made a posteriori, so that the grouping of classes in the sub-problems respects the class distribution in the input space. As another example, Shapire [18] combines ECOC with Boosting.

2 Decomposition

The decomposition scheme specifies the target function $f_l : \Omega \rightarrow \{-1, +1\}$ for subproblem l. To be valid, $\boldsymbol{f} = (f_1, \ldots, f_L)^\top$ should allow reconstruction, *i.e.* there should not be two pairs $(\boldsymbol{x}, k), (\boldsymbol{y}, k') \in T$, with $k \neq k'$ and $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{y})$. If no additional information is available, typically, all data of a same class will be associated to the same value by f_l. Therefore, the overall decomposition scheme can be specified by a *decomposition matrix* $\boldsymbol{D} \in \{+1, -1, 0\}^{L \times K}$ such that

$$D_{lk} = \begin{cases} +1 & \text{if class } \omega_k \text{ is associated with } +1 \text{ by } f_l \\ -1 & \text{if class } \omega_k \text{ is associated with } -1 \text{ by } f_l \\ 0 & \text{if class } \omega_k \text{ does not belong to the task of } f_l \ . \end{cases}$$

The validity of the decomposition scheme is expressed by the constraint that for every two columns of $\boldsymbol{D}$, there is at least one row for which the coefficients in the two columns are $+1$ and -1.

The subproblem l will be trained using all the information available, *i.e.* the training sample T_l used to learn f_l is the set of all the pairs $(\boldsymbol{x}, D_{lk})$ such that $D_{lk} \neq 0$ and $(\boldsymbol{x}, k) \in T$.

As illustration, the decomposition of the one-per-class and pairwise coupling schemes, for $K = 4$, are given by the decomposition matrices of Fig. 1 (a) and (b).

$$\begin{pmatrix} +1 & -1 & -1 & -1 \\ -1 & +1 & -1 & -1 \\ -1 & -1 & +1 & -1 \\ -1 & -1 & -1 & +1 \end{pmatrix} \qquad \begin{pmatrix} +1 & -1 & 0 & 0 \\ +1 & 0 & -1 & 0 \\ +1 & 0 & 0 & -1 \\ 0 & +1 & -1 & 0 \\ 0 & +1 & 0 & -1 \\ 0 & 0 & +1 & -1 \end{pmatrix}$$

(a) (b)

Figure 1: Classical decomposition matrices $\boldsymbol{D}$. Each row corresponds to one dichotomy and each column to one class. (a) illustrates the decomposition matrix of the one-per-class scheme; the matrix in (b) corresponds to the pairwise coupling scheme

3 Pairwise Coupling Reconstruction

In PWC, when an input vector $\boldsymbol{x}$ is to be classified, it is presented to all the classifiers, each providing a partial answer that concerns the two involved classes. Considering these answers as votes, a natural approach for the global classification consists in selecting the class that wins more votes. Assuming that the classifier discriminating between class ω_i (as positive) and class ω_j (as negative) computes an estimate $\hat{p}_{ij}$ of the probability

$$p_{ij} = P(\boldsymbol{x} \in \omega_i \mid \boldsymbol{x} \ , \ \boldsymbol{x} \in \omega_i \cup \omega_j) \ , \tag{1}$$

then, the classification is determined by

$$\arg \max_{1 \le i \le K} \sum_{j \ne i} [\![\hat{p}_{ij} > 0.5]\!] \ . \tag{2}$$

The operator $[\![.]\!]$ is defined as:

$$[\![\eta]\!] = \begin{cases} 1 & \text{if } \eta \text{ is true,} \\ 0 & \text{otherwise} \end{cases}$$

This combination considers the outputs of the classifiers as binary decisions. A different reconstruction approach consists in taking into consideration the fact that the outputs $\hat{p}_{ij}$ of the classifiers represent probabilities. Then, these values can be used to calculate approximations $\hat{p}_i$ of the a posteriori probabilities

$$p_i = P(\boldsymbol{x} \in \omega_i \mid \boldsymbol{x}) \ .$$

Considering a square matrix $\hat{\boldsymbol{P}}$ with the value $\hat{p}_{ij}$ in position $(i,j)_{i,j=1,\ldots,K,\ i \ne j}$ and with $\hat{p}_{ji} = 1 - \hat{p}_{ij}$, the values of the $\hat{p}_i$'s can be obtained from

$$\hat{p}_i = \frac{2}{K(K-1)} \sum_{j \ne i} \hat{p}_{ij} \ , \tag{3}$$

and the classification can then be given by

$$\arg \max_{1 \le i \le K} \ \hat{p}_i . \tag{4}$$

This procedure will hereafter be called *soft reconstruction*, as opposed to the voting procedure (2), referred to as *rough reconstruction*. The formulation given in (3) and (4) can be generalized to incorporate the two kinds of reconstruction:

$$\arg \max_{1 \le i \le K} \ \hat{p}_i \quad , \quad \hat{p}_i = \frac{2}{K(K-1)} \sum_{j \ne i} \sigma(\hat{p}_{ij}) \ ,$$

where σ takes the form of a threshold function at 0.5 for the rough reconstruction and the identity function for the soft reconstruction. Note that whenever σ is symmetric on [0,1], i.e. $\sigma(1-x) = 1-\sigma(x)$, then the p_i's sum to 1 and can thus be considered as probability estimates.

The two schemes presented explore the available information differently and about this a remark can be made. The following example follows from an observation made in [9]. Consider the matrix $\boldsymbol{P}$ with the $\hat{p}_{ij}$'s for a particular input vector $\boldsymbol{x}$ in a problem with three classes:

$$\begin{pmatrix} - & 0.6 & 0.6 \\ 0.4 & - & 0.9 \\ 0.4 & 0.1 & - \end{pmatrix} \tag{5}$$

It can be verified that $\boldsymbol{x}$ is classified as class ω_1 by a rough reconstruction, while a soft reconstruction will classify it as class ω_2. Given that these two functions may produce different outputs, some experiments were made to compare their performance. In addition, alternative forms for the function σ have been proposed and also experimented. The reconstruction schemes are summarized and labeled from PWC1 to PWC5 in Fig. 2.

$$\sigma(x) = \begin{cases} 1 & \text{if} \quad x \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \qquad \text{(PWC1)}$$

$$\sigma(x) = x \qquad \text{(PWC2)}$$

$$\sigma(x) = \frac{1}{1+e^{-12\,(x-0.5)}} \qquad \text{(PWC3)}$$

$$\sigma(x) = \begin{cases} 1 & \text{if} \quad x \geq 0.5 \\ x & \text{otherwise} \end{cases} \qquad \text{(PWC4)}$$

$$\sigma(x) = \begin{cases} x & \text{if} \quad x \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \qquad \text{(PWC5)}$$

Figure 2: The function σ used in the reconstruction schemes PWC1 to PWC5. The plots of these functions are presented in Fig. 3.

The schemes PWC1 and PWC2 are the threshold and linear combinations, while the sigmoidal function used in PWC3 is a compromise between the two. PWC4 is a semi-threshold function where the linear behavior of the first half range aims at recovering negative information that is close to the threshold and that might be caused by poor classifier performance. By negative information it is meant here the probability values below 0.5, which correspond to a negative vote in the case of a threshold function. PWC5 uses a concept similar to the previous scheme, but in this case negative information is given full meaning while positive information is given increasing importance with increasing distance from the threshold. Fig. 3 contains the plots of the different forms of σ.

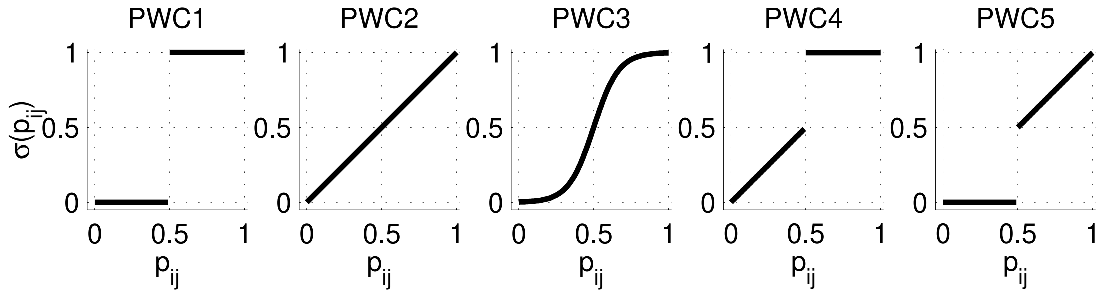

Figure 3: The different forms of the function σ as defined in Fig. 2.

In [13], a different way of approximating the class probabilities is proposed. Given

$$p_{ij} = \frac{p_i}{p_i + p_j}$$

and considering that for all i,

$$\sum_{j \neq i} (p_i + p_j) - (K-2)\, p_i = 1 \ ,$$

then the following expression can be derived:

$$p_i = \frac{1}{\sum_{j \neq i} \frac{1}{\hat{p}_{ij}} - (K-2)} \quad \text{(PWC6)} \ .$$

This scheme has been included in the experiments and is labeled PWC6.

The results of the experiments are presented in Fig. 4. The learning algorithm used to implement the classifiers is the decision tree algorithm C4.5 [15]. For each of the five databases from [12] that have been used in the tests, 20 runs with 3-folding were executed. The *3-fold cross-validated paired t test* [6] was applied to check for significant differences between the average error rates, with a confidence level of 0.95.

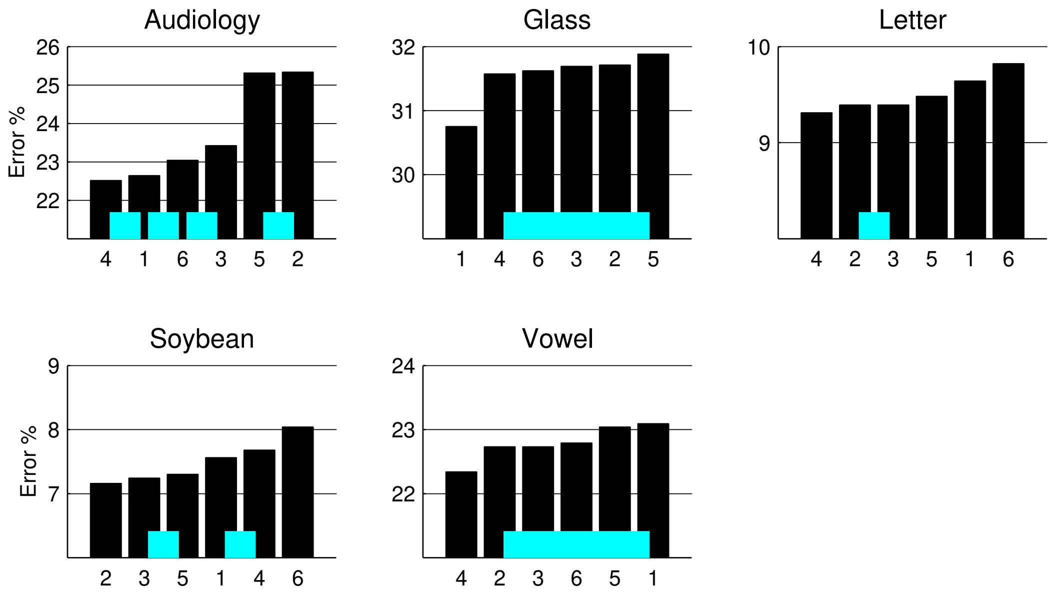

Figure 4: Comparison between different PWC reconstruction methods. Numerical values are presented in Table 1. Bars connected by the same light-coloured horizontal strip represent values that do not differ significantly (0.95).

The major outcome of Fig. 4 is that the results are quite regular, with none of the proposed reconstruction schemes performing significantly better than the others. This means that although the methods differ in behavior at a local level, they perform similarly globally, and thus there is no reason to give preference to any of them.

4 Improvement of the PWC Decomposition Scheme

A closer analysis of the PWC decomposition scheme reveals an important problem, which is related to the nonsense introduced by the values of $\hat{p}_{ij}$ when the item under consideration belongs neither to

Method	`audiology`	`glass`	`letter`	`soybean`	`vowel`
PWC1	22.64±4.2	30.75±4.6	9.64±0.4	7.56±1.5	23.09±2.2
PWC2	25.33±4.3	31.71±4.8	9.39±0.4	7.16±1.4	22.73±2.4
PWC3	23.42±4.1	31.69±4.6	9.39±0.4	7.24±1.4	22.73±2.4
PWC4	22.51±4.2	31.57±4.4	9.31±0.4	7.68±1.5	22.34±2.4
PWC5	25.31±4.3	31.88±4.9	9.48±0.4	7.30±1.5	23.04±2.2
PWC6	23.04±4.3	31.62±4.5	9.82±0.4	8.04±1.4	22.79±2.3

Table 1: Comparison between different PWC reconstruction methods. The values represent the average percentage of misclassification on the test set (and respective standard deviations).

class ω_i nor to class ω_j. Indeed, by (1), p_{ij} assumes that $\boldsymbol{x}$ is in class ω_i or in class ω_j, but for a given item $\boldsymbol{x}$, the estimation of the p_i's takes into account the outputs of all the pairwise classifiers, either significant or not.

For example, consider again the 3-class problem (5) in Section 3. If $\boldsymbol{x}$ belongs to class ω_1, p_{23} is absolutely irrelevant because the respective classifier has not been trained with data from class ω_1. Consequently, using it to find $\hat{p}$ is very likely to deteriorate the result of the calculation. If $\boldsymbol{x}$ does not belong to class ω_1, the high coefficient $p_{23} = 0.9$ is a strong indicator that $\boldsymbol{x}$ belongs to class ω_2, which will be selected by method PWC2. The problem is that the actual class of the item is obviously unknown a priori (this is precisely what we aim at determining) and thus the meaningful classifiers cannot be selected.

4.1 Correcting Classifiers

A procedure to overcome the referred problem is proposed here, which consists of, for each pairwise classifier separating class i from class j and with output $\hat{p}_{ij}$, training an additional classifier separating classes i and j from all the other classes, producing output $\hat{q}_{ij}$. The decomposition matrix notation is used in Fig. 5 to illustrate the case of four classes. The $\hat{q}_{ij}$'s provide an estimate that item $\boldsymbol{x}$ belongs to class i or class j, and it can be included in (3), which becomes:

$$\hat{p}_i = \frac{2}{K(K-1)} \sum_{j \neq i} \hat{p}_{ij} \cdot \hat{q}_{ij} \ .$$

The use of the correcting classifiers should cause the irrelevant $\hat{p}_{ij}$'s to loose significance and allow the quality of the estimation of the $\hat{p}_i$'s to be improved. This scheme is labeled PWC-CC.

$$\begin{pmatrix} +1 & -1 & 0 & 0 \\ +1 & 0 & -1 & 0 \\ +1 & 0 & 0 & -1 \\ 0 & +1 & -1 & 0 \\ 0 & +1 & 0 & -1 \\ 0 & 0 & +1 & -1 \end{pmatrix} \qquad \begin{pmatrix} +1 & +1 & -1 & -1 \\ +1 & -1 & +1 & -1 \\ +1 & -1 & -1 & +1 \\ -1 & +1 & +1 & -1 \\ -1 & +1 & -1 & +1 \\ -1 & -1 & +1 & +1 \end{pmatrix}$$

(a) Regular PWC matrix $\qquad$ (b) Additional classifiers

Figure 5: Decomposition matrices for the PWC-CC corrected scheme. (a) contains the standard PWC subproblems; the matrix in (b) corresponds to the correcting classifiers.

In the correcting scheme proposed, each value $\hat{q}_{ij}$ is in fact an estimate of $p_i + p_j$, which suggests a different approach for its calculation. In the OPC scheme referred in Sect. 1, the classifier of each subproblem provides directly an estimate $\hat{p}_i$. So, OPC can be used to perform the PWC correction task, by using the provided values of $\hat{p}_i$, that will be referred to as q_i, for all $i = 1,...,K$, to find the $\hat{q}_{ij}$'s. The great advantage of this combination is that the total number of classifiers is $\frac{1}{2}K(K+1)$, which is lower than $K(K-1)$ in PWC-CC, when $K > 3$. The label for the combination with OPC as correcting scheme is PWC-OPC.

Experiments were made to compare the performance of the corrected schemes with the standard ones. The same algorithm (C4.5), the same databases, and the same statistical test used in the experiments reported in Fig. 4 were used here. The standard PWC reconstruction scheme used is PWC2. Figure 6 contains the results; the performance obtained by C4.5 applied regularly as a multiclass algorithm is included for comparison.

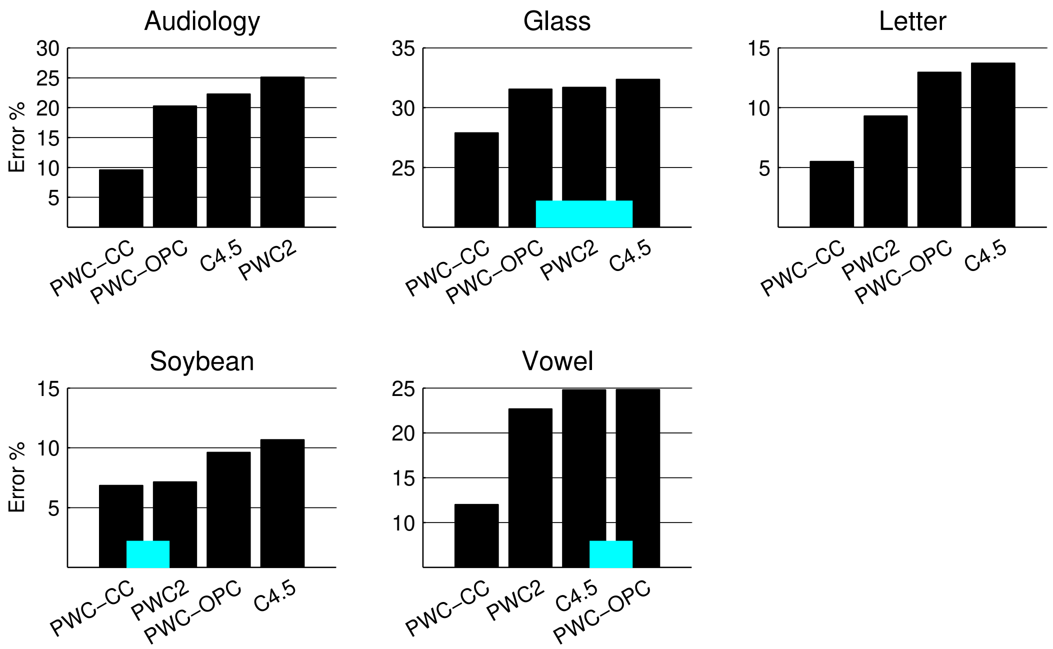

Figure 6: Comparison between standard and corrected PWC. Numerical values are presented in Table 2. Bars connected by the same light-coloured horizontal strip represent values that do not differ significantly (0.95).

Two important conclusions may be drawn from the experiments: 1) The correcting classifiers in PWC-CC introduce a significant correcting effect and improve the performance of PWC2 remarkably. 2) The performance of the PWC-OPC combination is worse than that of PWC-CC. For most of the databases, it even performs worse than PWC2.

4.2 Analysis of the Correcting Schemes Proposed

In order to investigate the cause of such a substantial difference between the performances of PWC-CC and PWC-OPC, experiments were made to compare the correcting component of each of the two combinations. Indeed, the set of additional classifiers used in PWC-CC can be used as a decomposition scheme by itself, as depicted in Fig. 5 (b), since it respects the constraint defined in Sect. 2. Its label is CC, by derivation. Given that the output of each classifier f_l is in $[0, 1]$ and that the decomposition matrix $\boldsymbol{D}$ is defined as in Fig. 5 (b) with values in $\{-1, +1\}$, the reconstruction in CC is made by:

$$\arg\max_k \sum_l (2f_l - 1) D_{lk} \ .$$

OPC is, as known, a decomposition scheme also. Figure 7 contains the results of the comparison; the performance of PWC2, PWC-CC, and C4.5 multiclass are included for reference. The same procedure was followed as in the experiments described previously.

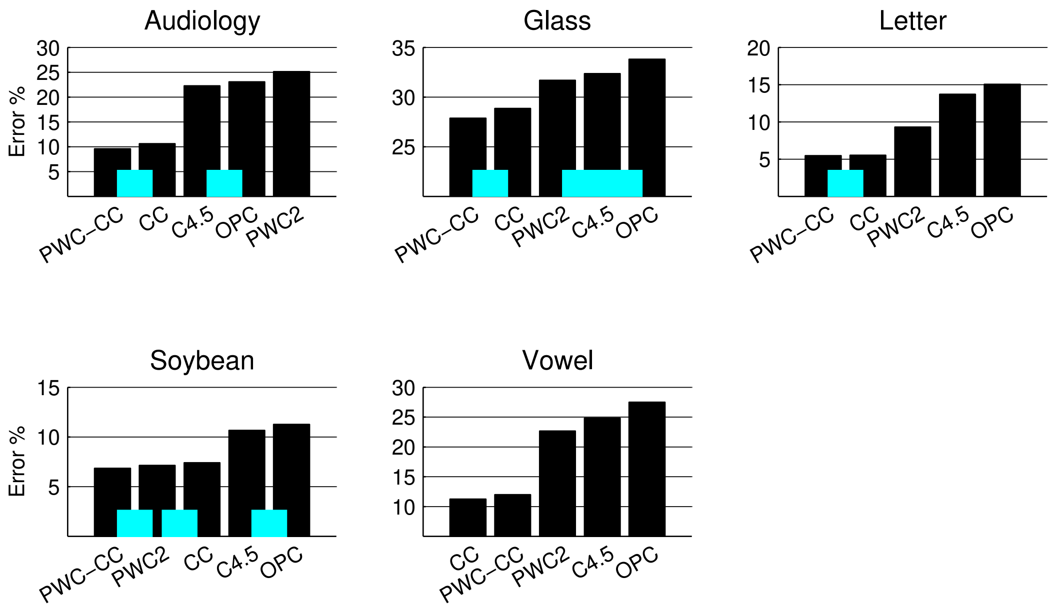

Figure 7: Comparison between the CC and the OPC correcting schemes when used as decomposition schemes. Numerical values are presented in Table 2. Bars connected by the same light-coloured horizontal strip represent values that do not differ significantly (0.95).

Method	audiology	glass	letter	soybean	vowel
C4.5	22.29±4.0	32.36±4.7	13.73±0.5	10.67±2.1	24.81±2.5
PWC2	25.11±4.1	31.69±4.8	9.31±0.4	7.15±1.4	22.68±2.4
PWC-CC	9.59±3.1	27.89±4.8	5.49±0.3	6.85±1.4	12.00±1.9
PWC-OPC	20.26±4.1	31.55±5.3	12.97±0.4	9.62±1.9	24.83±2.4
CC	10.62±2.6	28.87±4.4	5.54±0.3	7.40±1.4	11.24±1.9
OPC	23.06±3.5	33.80±5.5	15.04±0.4	11.26±2.0	27.49±2.4

Table 2: Comparison between the various decomposition schemes discussed in this paper. The values represent the average percentage of misclassification on the test set (and respective standard deviations).

From Fig. 7, it arises that the scheme CC has by itself a level of accuracy comparable to the combination PWC-CC, while OPC has a very poor performance. A logical explanation for this difference between the two schemes is their *class separability.* The class separability Δ of a decomposition scheme is defined as the minimal distance d_{cl} between every pair of columns (classes) in its decomposition matrix, and it has a major influence on the error recovering capability of the scheme. The distance measure d_{cl} between classes is defined as:

$$d_{\mathrm{cl}}(i,j) = |\{l \ : \ D_{li} \cdot D_{lj} = -1\}|.$$

For CC, $\Delta = 2(K-2)$, while $\Delta = 1$ in OPC. Considering that the reconstruction method adopted in OPC is the selection of the class associated with the classifier with the highest output, then, a single defective answer of any of the classifiers is likely to produce a misclassification. The reconstruction in CC allows, on the contrary, to recover from errors in the classifiers. In general, the error correcting capability of a decomposition scheme, when using a rough reconstruction, is at least

$$\left\lfloor \frac{\Delta - 1}{2} \right\rfloor .$$

As $\Delta = 2(K-2)$, CC allows the correction of at least $K-3$ errors.

The reasoning exposed above is not sufficient to explain the better performance of PWC-CC against PWC-OPC, because the correcting classifiers are not used directly in the reconstruction as normal classifiers, and thus they do not raise the class separability. The required explanation is, however, closely related to the one used to justify the difference between the correcting schemes alone and it also has to do with error-recovering capability. When OPC is used as correcting scheme, each output q_i will be used in the estimation of $K-1$ of the $\frac{1}{2}\,K(K-1)$ values of $\hat{q}_{ij}$. So, when one of the correcting classifiers makes a mistake, it will be propagated along part of the correcting scheme and subvert the global combination. In CC, each $\hat{q}_{ij}$ value is calculated by a single, different correcting classifier, which makes the global combination more tolerable to errors from the classifiers and thus more robust.

It can be concluded from Fig. 7 that the CC scheme may be preferred to standard PWC, with the disadvantage of having to train each classifier with all the data. This has the virtue, however, of eliminating the problem of the incompetent classifiers. Another interesting result is that CC is able to attain by itself the same level of performance as the combination PWC-CC in which it takes part. This result is in accordance with the theory behind the ECOC method, where the redundancy associated with a good class separation allows the reconstruction to be more robust, due to its error-correcting ability.

5 Conclusions

Decomposition by pairwise coupling is one of the existing techniques allowing a classification problem with several classes to be solved by a set of binary classifiers. It has been so far used in different applications, despite the fact that it suffers from the problem of using irrelevant information. That problem has been addressed here, and the proposed correcting procedure has been shown to be able to avoid the use of such information and to improve the performance of the decomposition scheme. Although that improvement is achieved at the cost of having twice as many subproblems as in standard PWC, with the additional fact that in the case of the correcting classifiers the whole training data is used to train each of them, it is also true that this task can be easily distributed, where in an ultimate solution all the classifiers can be trained and used in parallel.

As a side result, CC has been found to be a good decomposition scheme by itself. It is not affected by the problem of the incompetent classifiers and it can be favorably used as a replacement to standard PWC. The disadvantage is that all the data is used for the training of each classifier, which is negative from a point of view of training time.

Finally, it has been shown that the technique of decomposing by one-per-class has poor accuracy due to its high sensitivity to classifier performance.

References

[1] E. Boros, P. L. Hammer, Toshihide Ibaraki, A. Kogan, E. Mayoraz, and I. Muchnik. An implementation of logical analysis of data. RRR 22-96, RUTCOR–Rutgers University's Center For Operations Research, `http://rutcor.rutgers.edu:80/~rrr/`, Submitted, July 1996.

[2] L. Breiman, J. Olshen, and C. Stone. *Classification and Regression Trees.* Wadsworth International Group, 1984.

[3] Pierre J. Castellano, Stefan Slomka, and Sridha Sridharan. Telephone based speaker recognition using multiplt binary classifier and Gaussian mixture models. In *ICASSP*, volume 2, pages 1075–1078. IEEE Computer Society Press, 1997.

[4] Yves Crama, Peter L. Hammer, and Toshihide Ibaraki. Cause-effect relationships and partially defined boolean functions. *Annals of Operations Research*, 16:299–326, 1988.

[5] T. G. Dietterich and G. Bakiri. Error-correcting output codes : A general method for improving multiclass inductive learning programs. In *Proceedings of AAAI-91*, pages 572–577. AAAI Press / MIT Press, 1991.

[6] Thomas G. Dietterich. Statistical tests for comparing supervised classification learning algorithms. OR 97331, Department of Computer Science, Oregon State University,, 1996.

[7] Thomas G. Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.

[8] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, New York, 1973.

[9] Trevor Hastie and Robert Tibshirani. Classification by pairwise coupling. Technical report, Stanford University and University of Toronto, 1996. ftp://utstat.toronto.edu/pub/tibs/coupling.ps, to appear in the Proceedings of NIPS*97.

[10] E. B. Kong and T. G. Dietterich. Error-correcting output coding corrects bias and variance. In *The XII International Conference on Machine Learning*, pages 313–321, San Francisco, CA, 1995. Morgan Kaufmann.

[11] Eddy Mayoraz and Miguel Moreira. On the decomposition of polychotomies into dichotomies. In Douglas H. Fisher, editor, *The Fourteenth International Conference on Machine Learning*, pages 219–226, 1997.

[12] C. J. Merz and P. M. Murphy. UCI repository of machine learning databases. Machine-readable data repository `http://www.ics.uci.edu/~mlearn/mlrepository.html`, Irvine, CA: University of California, Department of Information and Computer Science, 1998.

[13] David Price, Stefan Knerr, Leon Personnaz, and Gerard Dreyfus. Pairwise neural network classifiers with probabilistic outputs. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems 7 (NIPS*94)*, volume 7, pages 1109–1116. The MIT Press, 1995.

[14] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[15] J. R. Quinlan. *C4.5 Programs for Machine Learning.* Morgan Kaufmann, 1993.

[16] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 63:386–408, 1958.

[17] Laszlo Rudasi and Stephen A. Zahorian. Text-independent talker indentification with neural networks. In *ICASSP*, volume 1, pages 389–392, 1991.

[18] Robert E. Shapire. Using output codes to boost multiclass learning problems. In Douglas H. Fisher, editor, *The Fourteenth International Conference on Machine Learning*, pages 313–321, 1997.

[19] P. Werbos. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences.* PhD thesis, Harvard University, 1974.

[20] Stephen A. Zahorian, Peter Silsbee, and Xihong Wang. Phone classification with segmental features and a binary-pair partitioned neural network classifier. In *ICASSP*, volume 2, pages 1011–1014. IEEE Computer Society Press, 1997.