International School on Neural Nets

"E.R. Caianiello”

7th Course

Ensemble Methods for Learning Machines

IIASS, 22-28 September 2002

Vietri sul Mare, Salerno - ITALY

Output Coding Decomposition Ensembles

Francesco Masulli

University of Pisa, Italy

masulli@disi.unige.it

Joint work with

Giorgio Valentini

University of Genoa, Italy

Outline

- Output Coding Decomposition Ensembles
- Open problems
- Experimental analysis of the factors affecting their effectiveness
- Application to Electronic Nose data
- Conclusions

Outline

- **Output Coding Decomposition Ensembles**
- Open problems
- Experimental analysis of the factors affecting their effectiveness
- Application to Electronic Nose data
- Conclusions

Ensemble Methods

- *Ensemble averaging*: linear combination of different learners (Perrone & Cooper, 1993; Hashemm, 1997);
- *Boosting & Bagging*: training set resampling (Freund & Shapire, 1996; Breiman, 1996);
- *Misture of experts*: non-linear combination of different learners (Jordan & Jacobs, 1994);
- *Feature selection*: learners based on groups of input features (Cherkauker, 1996);
- Etc.

Output Coding Decomposition Ensembles

Decomposition approach to classification:

• Splits a complex multiclass problem, or *polychotomy*, in a set of less complex and independent twoclass problems (*dichotomies*) and
•Recomposes the outputs of dichotomizers, in order to solve the original polychotomy .

Learning machines composed by two main units:
•**Decomposition Unit** that analyzes the input pattern and calculates the codeword using an assigned decomposition scheme.
•**Decision Unit** that associates the computed codeword with a class.

Decomposition approach to classification

•Splits a complex multiclass problem, or *polychotomy*, in a set of less complex and independent twoclass problems (*dichotomies*) and

•Recomposes the outputs of dichotomizers, in order to solve the original polychotomy .

Learning machines composed by two main units:

•Decomposition Unit that analyzes the input pattern and calculates the codeword using an assigned decomposition scheme.

•Decision Unit that associates the computed codeword with a class.

Voting and decomposition approach to classification

• ***Homogeneous voting*** (e.g., Perrone,1993; Meir,1994; Breiman, 1994): Multiple runs of the same algorithm on the same learning problem are combined by voting. It can only *reduces variance*.

• ***Non- homogeneous voting*** (e.g., Shapire,1990; Quinlan, 1993b): Voting multiple hypotheses constructed by different learning algorithms applied to the same problem. It can *reduce both bias and variance* if the various algorithms are different.

Decoding (reconstruction) of a codeword in the decomposition approach to classification *is equivalent to a vote* among those dichotomizers that learned the relevant boundaries (Kong&Dietterich,1995)

Voting only improve performances if the errors made by various voters are not “highly” correlated.

Decomposition Unit

Let be

- X multidimensional space of attributes
- $C_1,...,C_K$ labels of classes.
- $P : X \to \{C_1,...,C_K\}$ K- classes polychotomy (or *K-polychotomy*),

The decomposition of P generates a set of L dichotomizers $f_1,..., f_L$

A dichotomizer f_i is a discriminating function that subdivides C_i^+ and C_i^-, the input patterns in two disjoint *superclasses* each of them grouping a subset of classes of the *K-polychotomy*

Decomposition Matrix

$$D = \left[d_{ik}\right] \qquad i = 1,\ldots,I \quad k = 1,\ldots,K$$

represents the decomposition and

connects classes $C_1,\ldots,C_K$ to the superclasses C_i^+ and C_i^-

$$d_{ik} = \begin{cases} +1 \ \text{ if } \ C_k \subset C_i^+ \\ -1 \ \text{ if } \ C_k \subset C_i^- \\ \ \ 0 \ \text{ if } \ C_k \cup (C_i^+ \cup C_i^-) = \emptyset \end{cases}$$

Representations of a Decomposition Matrix

Class codewords

		c_1	c_2	c_3	c_4
Dichotomies	f_1	+1	-1	0	-1
	f_2	+1	0	-1	+1
	f_3	-1	-1	+1	0
	f_4	+1	0	+1	+1
	f_5	+1	+1	0	1
	f_6	+1	-1	-1	+1
	f_7	-1	0	+1	0

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

Reconstruction Unit

In this stage, a pattern is assigned to the class whose codeword is most similar to the output of the set of dichotomizers.

$$class = \arg \underset{k}{\vee} sim(F, c_k)$$

where :

$F = (f_1, \ldots, f_I)$

c_k codeword of class C_k

sim similarity measure

Dichotomizers outputs	*sim*
discrete	Hamming distance or similar
continuous	Inner product, L1 or L2 norms

Decomposition Schemes (DS)

A DS decomposes a polychotomy into a set of dichotomies

- ***A priori decomposition schemes***:
  - **One-Per-Class** (OPC) *(Nilsson, 1965)*
  - *Minimal* (MIN) *(Moreira, Mayoraz, 1997)*
  - *Maximal* (MAX) *(Moreira, Mayoraz, 1997)*
  - **Output Distributed Codes** (ODC) *(Sejnowski, Rosenberg, 1987)*
  - **PairWise Coupling** (PWC) *(Hastie, Tibshirani 1996)*
  - PairWise **Correcting Classifiers** (CC) *(Moreira, Mayoraz, 1998)*
  - **Error Correcting Output Codes** (ECOC) *(Dietterich, Bakiri, 1991, 1995)*
- ***A posteriori decomposition schemes*** *(Mayoraz, Moreira, 1996)*

One-Per-Class DS

Classical approach (Nilsson,1965)

Each dichotomy separates a single class from all others

K classes ⇒ K dichotomies

OPC decomposition matrix (4 classes)

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

Decision boundaries between couple of classes are learned only twice

Minimal DS

(Mayoraz&Moreira,1997)

K classes ⇒ $I = \lceil \log_2(K) \rceil$ dichotomies

MIN decomposition matrix (4 classes)

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

Decision boundaries between couple of classes are learned once

Maximal DS

(Mayoraz&Moreira,1997)

All possible dichotomies

K classes $\Rightarrow \frac{1}{2}(3^K + 1) - 2^K$ redundant dichotomies.

We delete:

- equivalent dichotomies like $f^{'} = -f$
- trivial dichotomies like $f^{-1}(-1) = \emptyset$

K classes $\Rightarrow$ $I = 2^{K-1} - 1$ useful (i.e., not redundant) dichotomies

MAX decomposition matrix (4 classes)

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

PairWise Coupling DS

(Moreira&Mayoraz,1998)

Each dichotomy separates a class c_i from class c_j ignoring all other classes

K classes ⇒ $I = \binom{K}{2}$ dichotomies

PWC decomposition matrix (4 classes)

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

Variants of PWC DS:

Correcting Classifiers (CC) decomposition scheme

CC decomposition

matrix (4 classes)

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

Correcting Classifiers (PWC-CC) decomposition scheme

Error Correcting Output Codes DS (ECOC)

(Dietterich&Bakiri, 1991,1995)

Coding theory ⇒ classification problems

Large decomposition schemes based on ECOC as class codewords:

- redundancy of codewords gives error recovering capabilities to the reconstruction unit ⇒ An ECOC DS allows a correct classification even if a subset of dichotomizers are wrong.
- decision boundaries between pairs of classes are learned many times

Error Correcting Output Codes DS

ECOC decomposition problem (4 classes)

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

Decision boundaries between pairs of classes are learned many times

ECOC effectiveness (1)

The maximal number of *errors that can be corrected* in an ECOC DS is

$$MaxNE = \left\lfloor \frac{\Delta_D - 1}{2} \right\rfloor$$

where Δ_D is the minimal Hamming distance (MDH) between pairs of columns (codewords) in the decomposition matrix **D.**

$\Rightarrow$ *Column separation* - a codeword must be far from the other codewords of the decomposition matrix (Hamming distance).

ECOC effectiveness (2)

ECOC are effective if errors induced by channel noise on single bits are independent (Peterson,1972).

If an ECOC DS contains **very similar rows** (dichotomies) each error of an assigned dichotomizer will be likely to appear in the most similar dichotomizers.

$\Rightarrow$ *Row separation* - dichotomizers f_i and f_j $\forall i \neq j$ should be not correlated $\Rightarrow$ each row should be far from the other rows and from their complements (Hamming distance).

ECOC generation algorithms

- Exhaustive algorithm (MAX Decomposition)
- Bose Chauduri Hocquenghem (BCH) algorithm (1960,1959)
- Random climbing up algorithm (Dietterich&Bakiri,1995)
- Random Codes (RC) *(James, 1998)*

•Constrained random codes

ECOC Exhaustive algorithm

- given a Hamming distance, it maximizes the distance among codewords
- equidistance between couple of codewords
- “Bayes consistent” (James,1998): If each dicothomizer approximates the Bayes (optimal) discriminant function then the overall polychotomizer will produce Bayes Classification
- Problem: exponential growth of the dichotomies with the number of classes

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

ECOC exhaustive decomposition matrix (4 classes)

Bose Chauduri Hocquenghem (BCH) algorithm - 1

- **Algebraic method** developed by Bose, Chauduri and Hocquenghem based on polinomial representation of the finite Galois fields (Bose, Chauduri, 1960).
- The maximization of the Hamming distance for a fixed codeword length is in general suboptimal. **BCH ECOC are not Bayes optimal**, but allow to generate ECOC **codewords of tractable length**
- The algorithm was originally employed for error recovering in serial transmission of data

Bose Chauduri Hocquenghem (BCH) algorithm - 2

- *Problem*: the algorithm try to maximixe the distance among codewords, but can duplicate rows of the decomposition matrix or generate trivial dichotomies
- In the context of classification problems the *correlations among codeword bits become significant*.
- We have a bit *modified the original algorithm*, testing also the distance among rows of the generated ECOC decomposition matrix: Rows identical or below a desired Hamming distance are deleted.

Recent development in ECOC machines

- Combination of ECOC and Boosting techniques (Shapire, 1997);
- Finding a DS minimizing the Empirical Loss is NP complete (Crammerr & Singer, 2000) ➔ DS with continuous codes in order to make the problem tractable using a constrained quadratic optimisation problem.

Recent improvements to ECOC DS

- Random selected DS => well-separated codewords (Berger, 1999);
- Circular ECOC ➔ reduce sensivity to codeword selection (Ghaderi & Windeatt, 2000);
- Binary labelling techniques ➔ reduces the correlation between base learners (Windeatt $ Ghaderi, 2001)

Successful applications of Output Coding Decomposition Ensembles

Improvements over standard k-way classifiers

- Classification of cloud types (Aha & Bankert, 1997);
- Text classification (Berger, 1999; |Ghani, 2000);
- |Food qualification (Pardo, Sberveglieri, Masulli & Valentini, 2001);
- Face verification (Kittler, Ghaderi, Windeatt & Mathas, 2001);
- Bioinformatics (Valentini, 2002).

Why Output Coding Decomposition Ensembles generalize so well?

- Reduction of both bias and variance (Kong &Dietterich, 2000; |Berger, 1999)
- Large margin classifiers framework (Shapire, Freund, Bartlett & Lee, 1998; Allwein, Shapire & Singer, 2000).

Outline

- Output Coding Decomposition Ensembles
- **Open problems**
- Experimental analysis of factors affecting their effectiveness
- Application to Electronic Nose data
- Conclusions

Output Coding Decomposition Ensembles
OPEN PROBLEMS - 1

- **Experimental analysis of the trade-off between error recovering capabilities and learnability of the dichotomies induced by the decomposition scheme**. Theoretical analyses: Allwein, Shapire & Singer (2000).
- **Study of the relationship between codeword length and performances**. Preliminary results: Ghani (2000).
- **Selection of optimal dichotomizers** for the DU. Addressed by: Berger (1999), Ghani (2000), Masulli & Valentini (2000).

Output Coding Decomposition Ensembles OPEN PROBLEMS - 2

- How to design codes jointly maximizing the distance between rows and columns of the DM **(a-priori methods**).
- How to design codes for a given multiclass problem (**a-posterior methods**)
  - Greedy approach (Mayoraz & Moreira, 1997)
  - Soft weight sharing (Alpaydin & Mayoraz, 1999)
  - Continuous codes & constrained optimisation problem (Crammer & Singer, 2000)
- How to relate **performances of ECOC and dependence among output errors** (Kong & Dietterich, 1995; Guruswami & Sahai, 1999).

Learning Machines implementing the Decomposition Unit

Implementation of decomposition schemes to classification:

• **monolithic classifier**: MIMO learning machine (e.g, MLP, Decision trees, etc.) trained on the full training set to produce the right codewords on its outputs. We say that each output of a monolithic classifier is an implicit dichotomizer.

• **parallel classifiers**: or ***Output Coding Decomposition Ensembles:*** *L* independent dichotomizers (e.g., Simple Perceptrons, Support Vector Machines, and, again, MISO MLP, Decision trees, etc.) each one trained independently on a specific dichotomic tasks using the full training set.

Parallel Classifiers

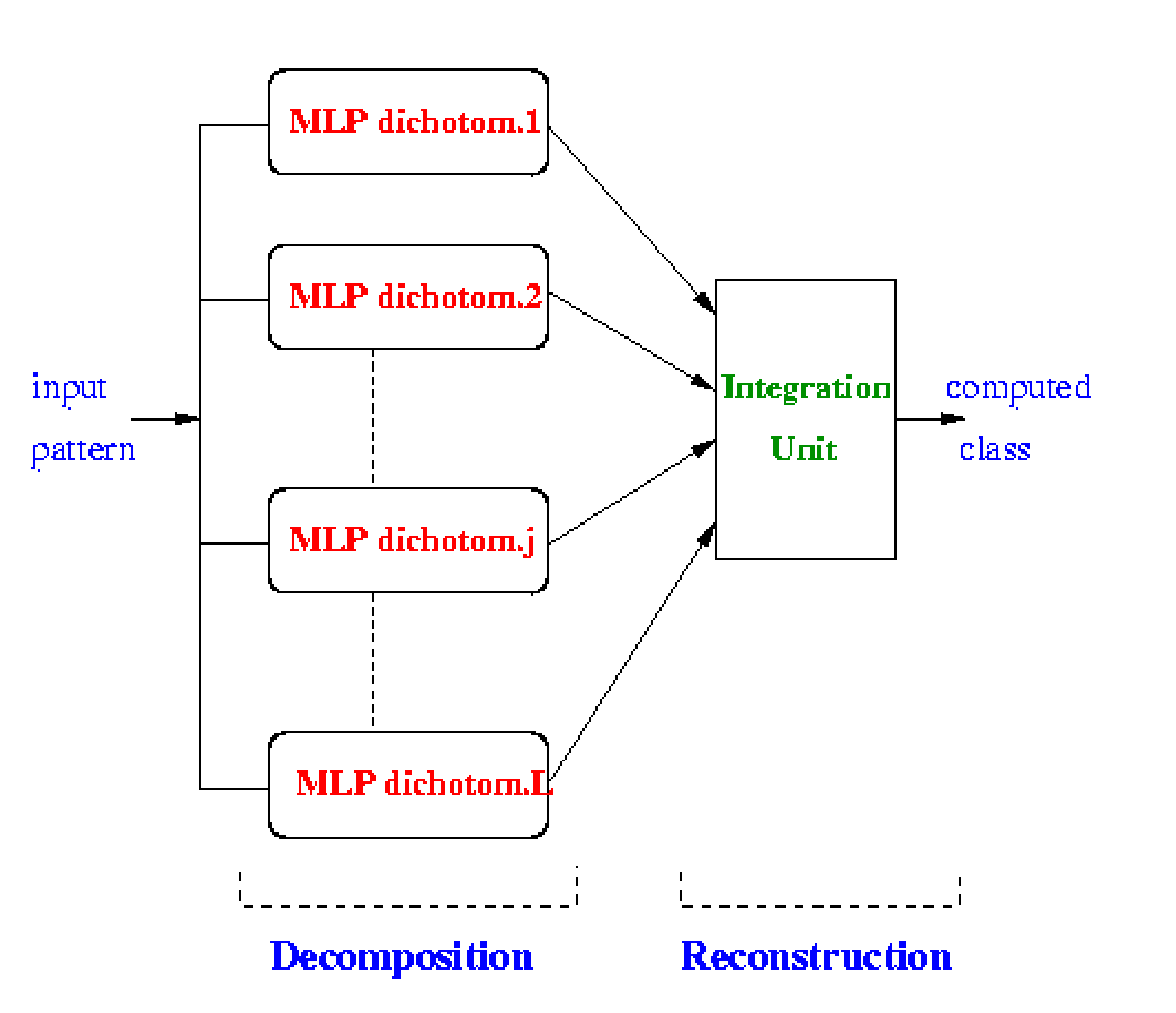
Parallel Linear Classifiers (PLD)
(Alpaydin&Mayoraz,1998)

Parallel multiclassifiers based on decomposition of polychotomies into dichotomies using a separate *linear learning machine* for implementing each dichotomizer.

Parallel Non-linear Classifiers (PND)

Parallel multiclassifiers based on decomposition of polychotomies into dichotomies using a separate *non-linear learning machine* for implementing each dichotomizer.

PND using MLP dichotomizers



Outline

- Output Coding Decomposition Ensembles
- Open problems
- **Experimental analysis of factors affecting their effectiveness**
- Application to Electronic Nose data
- Conclusions

Problems we address now

Experimental analysis of **factors affecting the effectiveness of ECOC methods**.

In particular we focus on the following items:

- *Architecture of the decomposition unit.*
- *Dependency among codeword bits coding the classes.*
- *Decoding function selected for the decision unit.*
- *Relationships between ensemble accuracy, base learner accuracy and error correcting power.*

Data sets

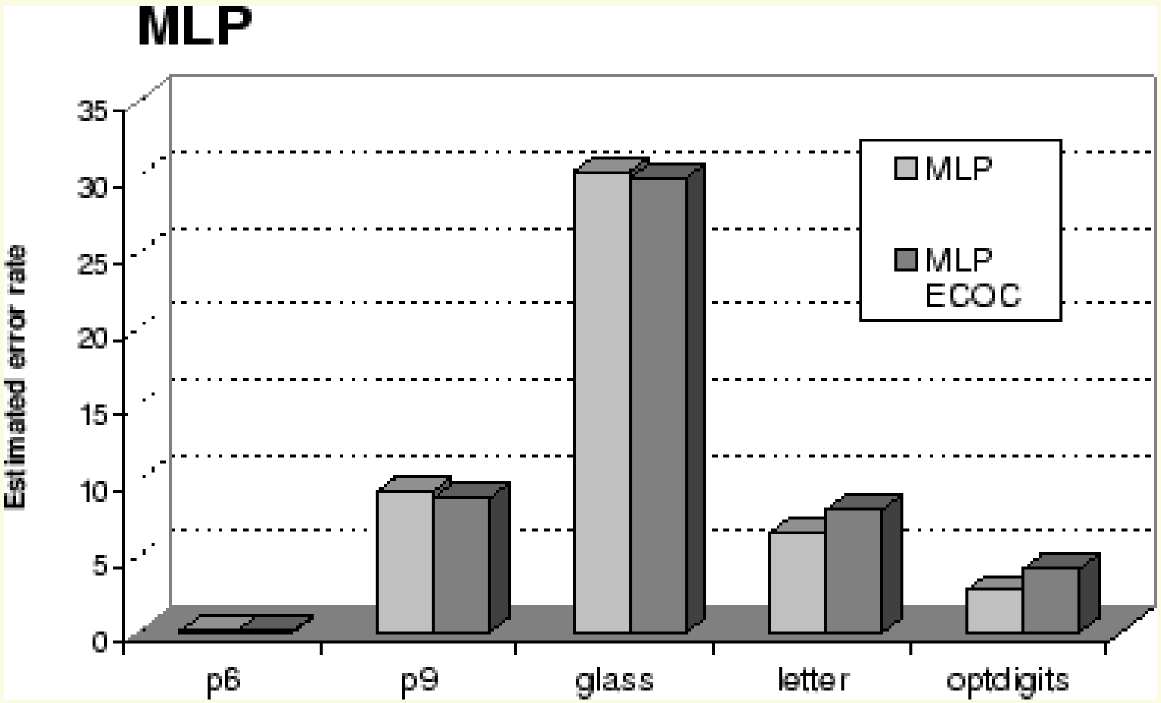
Data sets	**# attributes**	**# classes**	**#training samples**	**# testing samples**
d5	3	5	30000	30000
p6	3	6	1200	1200
p9	5	9	1800	5-fold cross-val
glass	9	6	214	10-fold cross-val
letter	16	26	16000	4000
optdigits	64	10	3823	1797

p6 / p9 synthetic - normal distributed clusters of data classes without/with overlaps.

d5 synthetic – each class 2 disjoint gaussian clusters

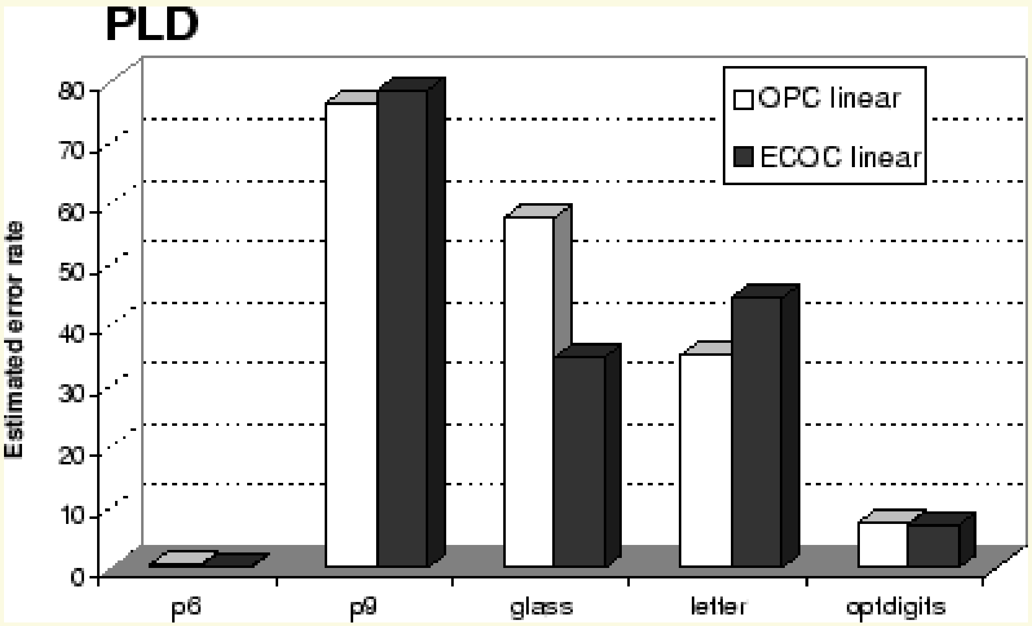
Glass, *letter* and *optdigits* from *UCI repository*.

Architectures – Results - 1



ECOC MLP monolithic classifiers do not outperform standard MLP (consistent with Dietterich & Bakiri, 1995)

Architectures – Results - 2

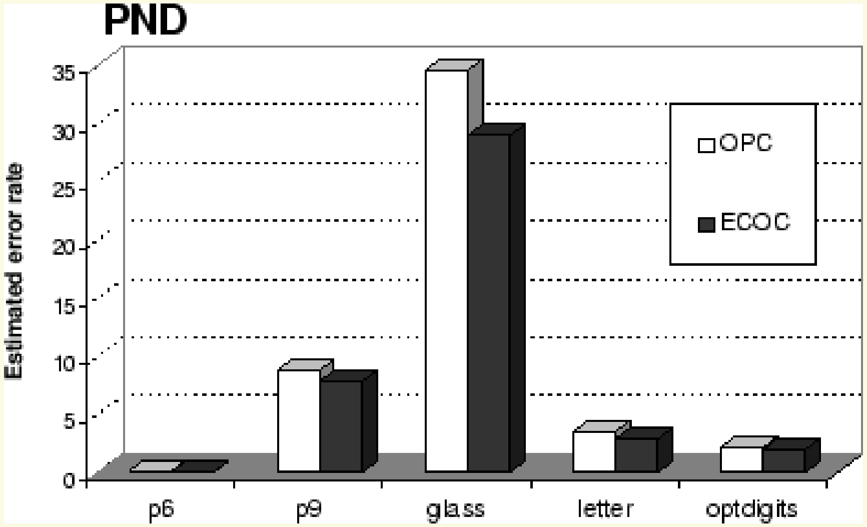


•**data sets p6, p9, and optdigits** no significant statistical difference among OPC and ECOC decomposition,

•**glass** data PLD ECOC outperforms all other types of polychotomizers

•**letter** PLD OPC achieve better results.

Architectures – Results - 3



•**data sets p6 and optdigits**: no significant differences among OPC and ECOC PND can be noticed.
•**p9 data set**, ECOC shows expected errors significantly smaller than OPC.
•**glass and letter data sets** expected errors are significantly smaller for ECOC compared with OPC.

Architectures – Results - 4

ECOC PND show expected error rates significantly lower than OPC PND.

PLD show remarkable higher errors over all data sets, and in particular they fail over *p9*.

Summarizing:

- **the expected errors are significantly smaller for *PND* compared with direct monolithic MLP classifiers and *PLD***
- **ECOC outperforms OPC decomposition only in *PND* ensembles.**

Architectures – Discussion - 1

Question: *Why PND perform better than ECOC monolithic learning machines?*

- PND dichotomizers are less complex than ECOC *monolithic* learning machines ➔ better generalization capabilities.
- **In PND each codeword bit is learned and computed by its own MLP**, specialized for its particular dichotomy, while in **monolithic classifiers** each codeword bit is learned and computed by a (non) linear combination of **hidden layer** outputs ➔ higher correlation among codeword bits

Architectures – Discussion - 2

Question: *Why PND perform better than PLD learning machines?*

In PLD the error recovering capabilities induced by ECOC are counter-balanced by higher error rates of linear dichotomizers.

Dependency ←→Effectiveness

Code theory: if errors on different codeword bits are dependent, the s of error correcting code is reduced (Peterson and Weldon, 1972) →**Dependence among output errors affects the effectiveness of ECOC methods** (Kong & Dietterich, 1995; Guruswami & Sahai, 1999; Masulli & Valentini, 2000)

Levels:

- **Codeword level**: if a DM contains very similar rows (dichotomies), each error of an assigned dichotomizer will be likely to appear in the most correlated dichotomizers →reduction of effectiveness of ECOC.
- **Architectural level**

Measurement of Dependence among Output Errors - 1

Masulli & Valentini (2001) proposed some measures of dependence of output errors based on **mutual information**.

Mutual information (MI) measures the matching between the joint probability density distribution and the product of the marginal probability density distribution of the output errors

MI special case of the Kullback-Leibler divergence between two distributions

Measurement of Dependence among Output Errors - 2

Def: mutual information error index

$$\Phi_E = \sum_{i=1}^{l} \sum_{j=1}^{l} I_E(e_i, e_j)$$

it is the sum of the mutual information of the output errors between all the output pairs of the learning machines

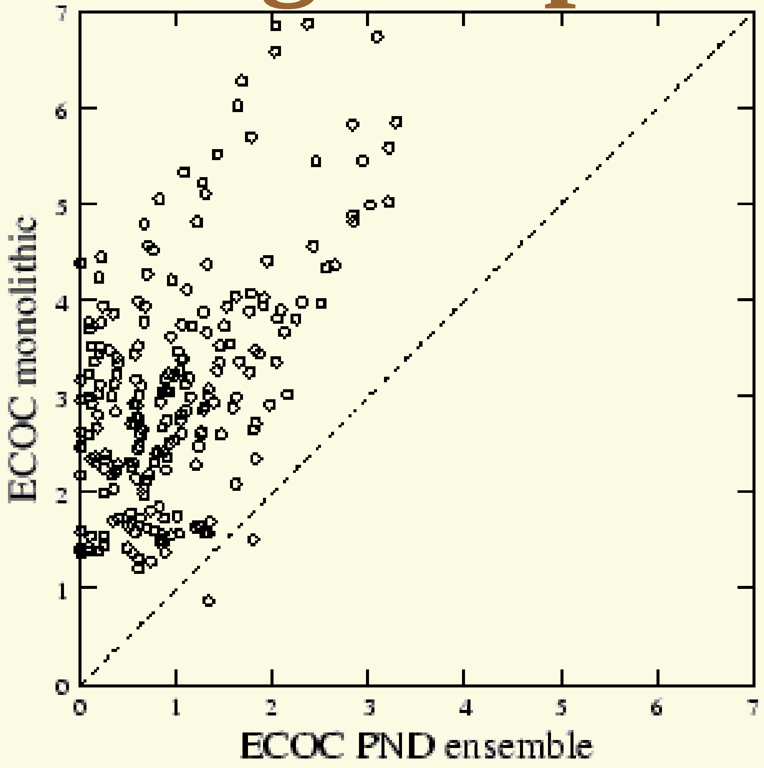
➔computable quantity to estimate the dependence between codeword bit errors

An high value of Φ_E *corresponds to an high dependence between output errors and vice versa.*

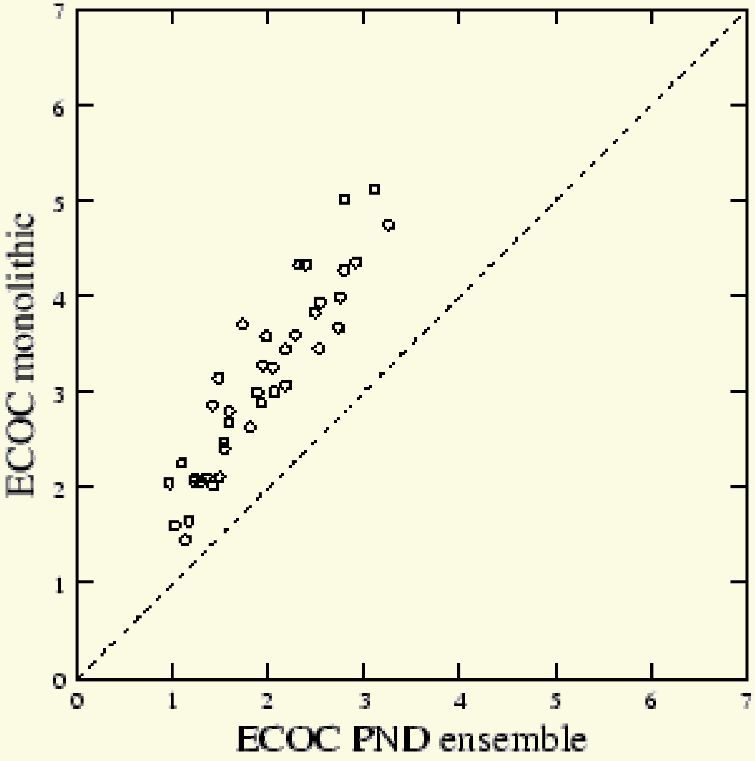
Measurement of Dependence among Output Errors - 4

- Each point corresponds to ECOC learning machines implemented with MLP with different number of hidden units and using different partitions of the output error.
- On all the data sets about all the points are above the dotted line, i.e. all the values of δ_{E} are greater for ECOC *monolithic* compared with ECOC *PND*.
- The results show that ***monolithic*** **architectures are affected by a higher dependence among codeword bit errors**. *This is* ***consistent*** *with the previous discussion, at architectural level, about the interdependence among monolithic MLP ECOC outputs.*

Measurement of Dependence among Output Errors - 5

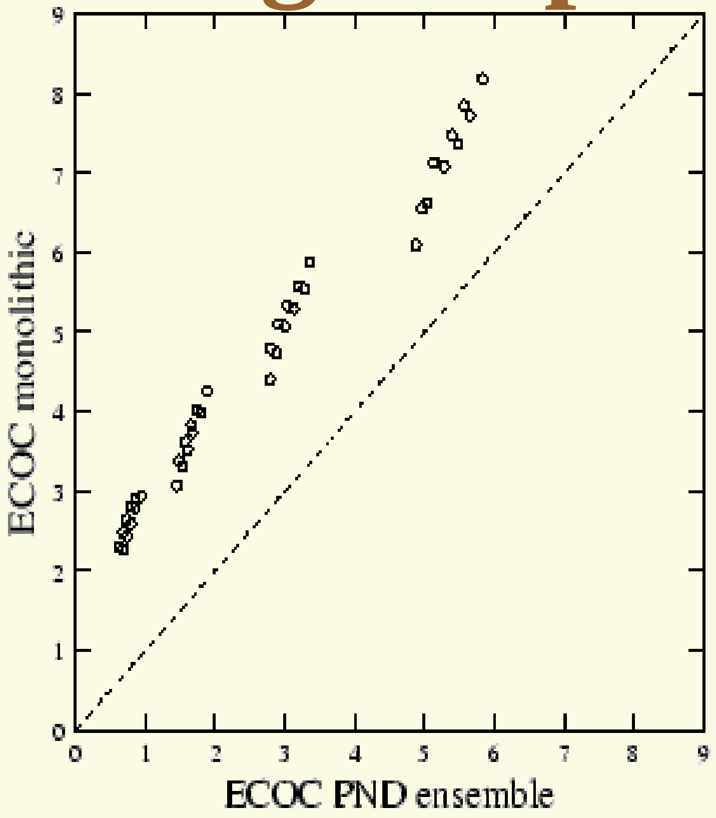


Data set: *d5*

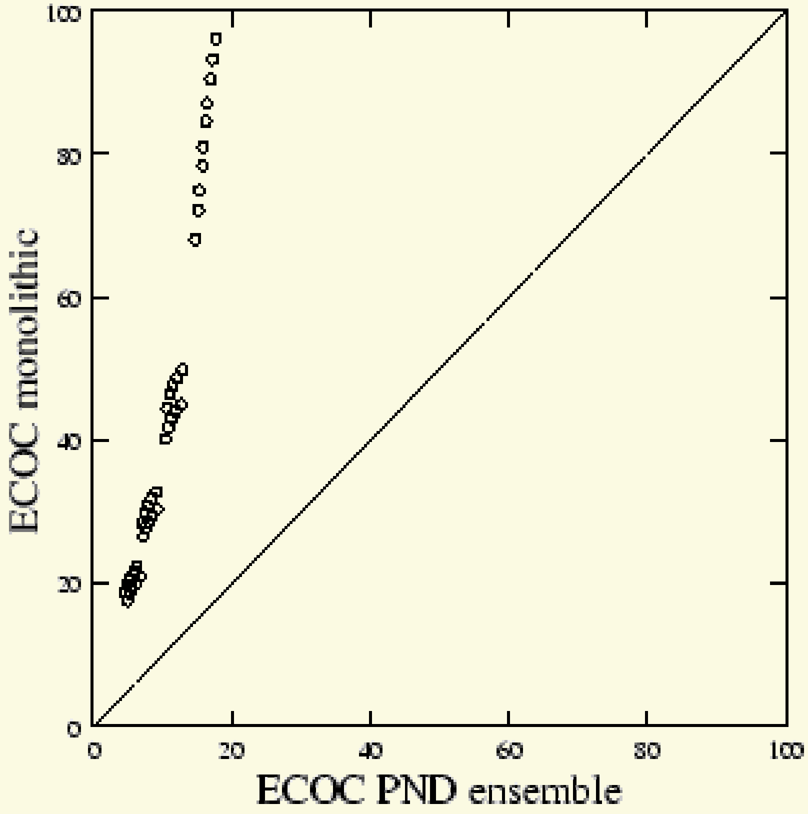


Data set: *glass*

Measurement of Dependence among Output Errors - 6



Data set: *opdigits*



Data set: *letter*

Ensemble Accuracy and Decoding Function - 1

- *Can the choice of a particular decoding function affects the performance of ECOC MLP ensembles*?
- *How the minimum Hamming distance with fixed length codewords affects the effectiveness of an ECOC ensemble?*

Ensemble Accuracy and Decoding Function - 3

Decoding functions based on

- Hamming distance: $$\mathcal{D}_{Hamm}(x) = \arg\min_j \sum_{i=1}^{n} \frac{1}{2} |D_{ij} - C_i(x)|$$
  if $C_i(x) \in \{-1, +1\}$
- L1 norm: $$\mathcal{D}_{L_1}(x) = \arg\min_j \sum_{i=1}^{n} |D_{ij} - C_i(x)|$$
  if $C_i(x) \in \mathbb{R}^d$
- L2 norm: $$\mathcal{D}_{L_2}(x) = \arg\min_j \sum_{i=1}^{n} (D_{ij} - C_i(x))^2$$
  if $C_i(x) \in \mathbb{R}^d$

Ensemble Accuracy and Decoding Function – Methods 1

We generated ECOC decomposition matrices with *constrained random algorithms.*

Constraints in order to

- eliminate
  - trivial dichotomies (e.g. rows with all +1 or all -1), and
  - equal or complementary rows (i.e. dichotomies)
- achieve a desired minimum Hamming distance (MHD) between the columns (codewords) of the DM.

Ensemble Accuracy and Decoding Function – Methods 2

- Generalization error estimated using 5-fold cross validation.
- Data set 1, 2: *optdigits* and *image-segmentation* data sets from the UCI repository (merge of training and test sets).
- Data set 3: synthetic *p20* data set (NEURObjects). 20 classes 3-dimensional - each class 3 disjoint clusters of data normally distributed with diagonal covariance matrices.

Ensemble Accuracy and Decoding Function – Results 1

Considering the proposed data sets:

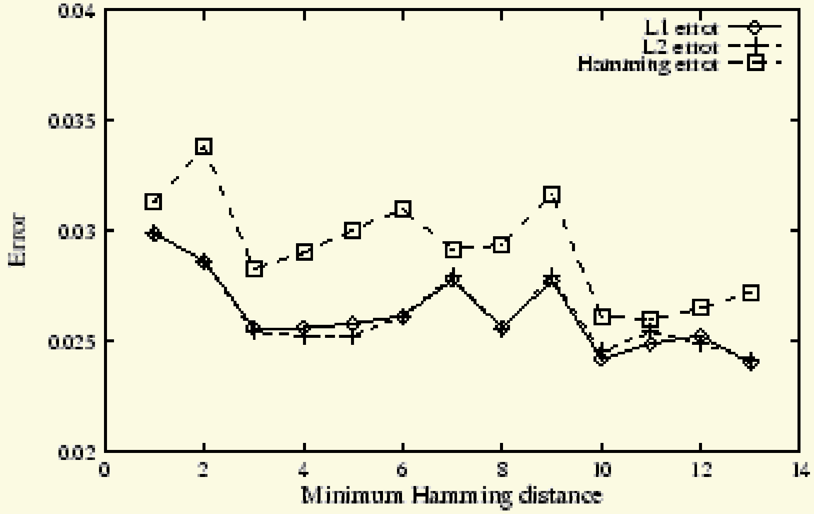
- The comparison of the estimated generalization error of the ECOC ensembles shows that the **decoding functions based on the L1 and L2 norms outperform the decoding based on the Hamming distance**.

  Only on the *p20* data set with 50-bit ECOC ensembles with linear perceptrons as base learners there is no difference between L1 norm and Hamming distance based decoding, but in this task the ECOC ensemble clearly fails, performing a sort of random guessing.

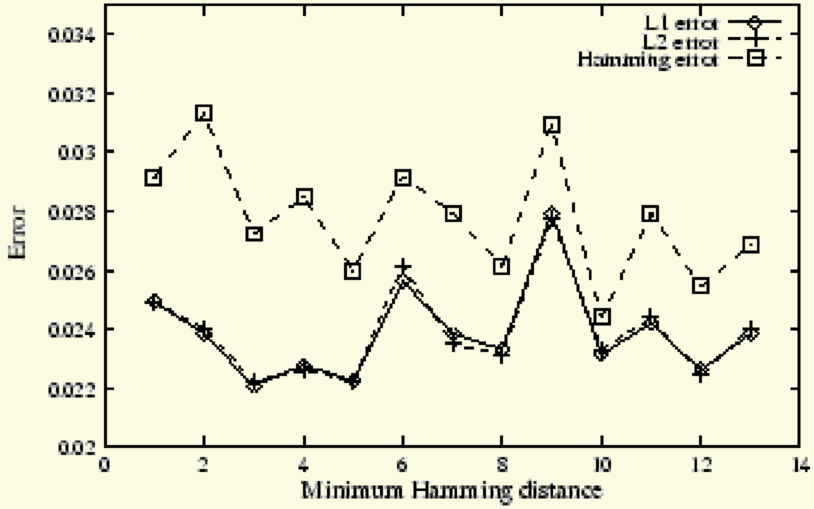
- there is **no significant difference using L1 or L2** norms in decoding

Ensemble Accuracy and Decoding Function – Results 2

optdigits data set



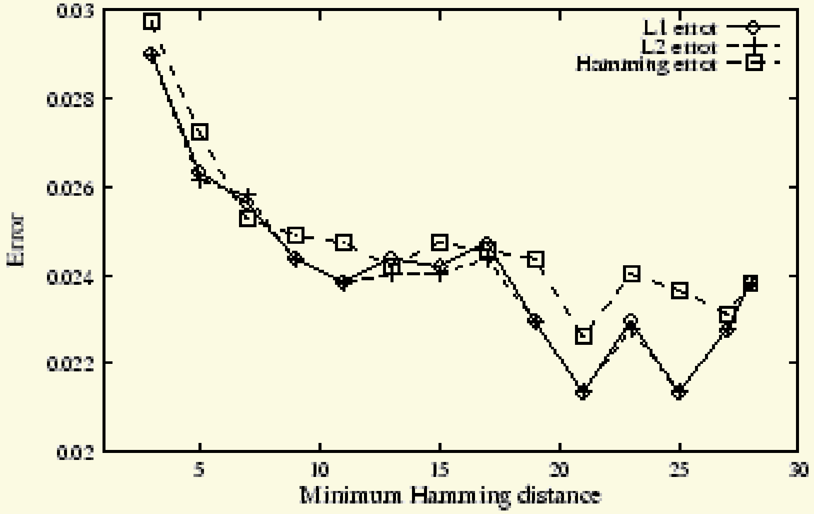
32 bits ECOC
MLP 4 hidden



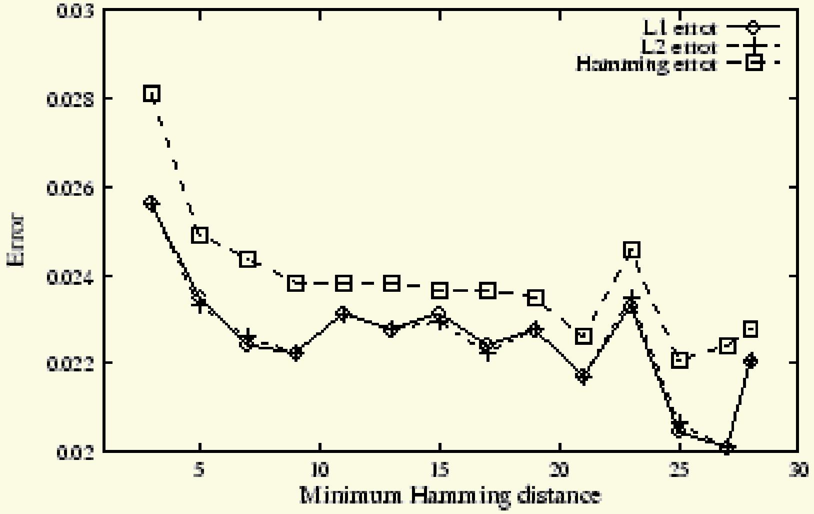
32 bits ECOC
MLP 10 hidden

Ensemble Accuracy and Decoding Function – Results 3

optdigits data set



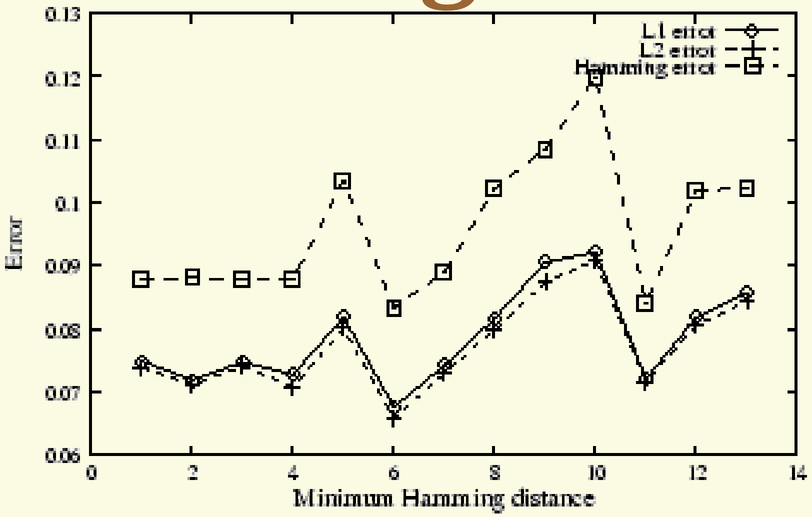
32 bits ECOC
linear perceptron



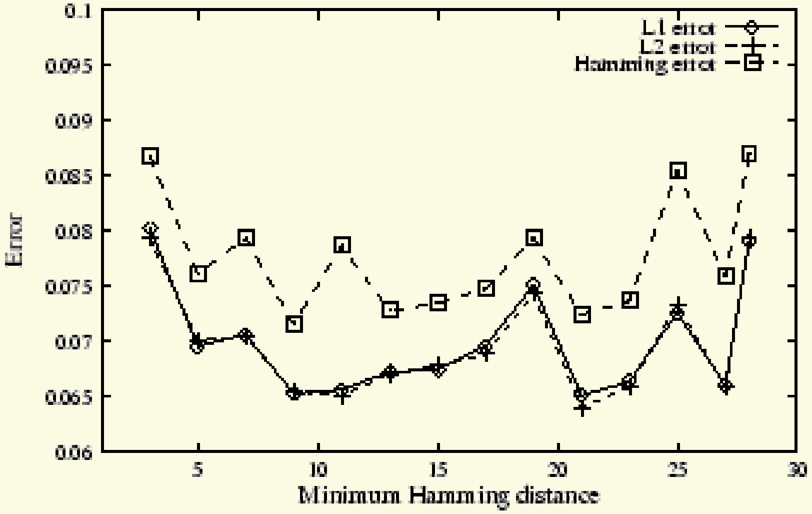
64 bits ECOC
linear perceptron

Ensemble Accuracy and Decoding Function – Results 4

optdigits data set



32 bit ECOC
MLP 4 hidden



32 bit ECOC
MLP 10 hidden

Ensemble Accuracy and Decoding Function – Discussion

Decoding function plays an important role

- *L1* and *L2* norm distance seem to be well-suited for the decoding
- Hamming distance based decoding function achieves worse results

***L1* and *L2* norms exploit the "confidence"** in the prevision of each base learner, while **Hamming decoding** discards all the information except the **hard membership** to a class

Generalization error & MHD

- Relationship between the **estimated generalization error and the minimum Hamming distance (MHD)** between the codewords.

Expected a monotonic decrement with the MHD

not confirmed: trends not regular and only in a few cases (shown in previous slide) we can observe a monotonic decrement of the error with MHD.

ECOC ensemble performance vs. base learner accuracy - 1

Question: *Why the selection of different base learners affects in a so significant manner the performance of the ensemble?*

➔Experimental analysis of relationship among:

- overall ensemble error
- average base learner error
- minimum Hamming distance (MHD) between the codewords.

Note: *The error recovering power of ECOC methods depends on the MHD between codewords (if the output errors of the decomposition unit are independent)*

ECOC ensemble performance vs. base learner accuracy - 2

Experimental results:

- *optdigits* data set:
  - average base learner error increments with MHD minimum Hamming distance
  - ensemble error tends to decrease with the MHD, especially using long codewords
- *image-segmentation* data set:
  - average base learner error increases with MHD
  - ensemble error oscillates around .035 and .040
- *p20* synthetic data set:
  - average base learner error not clear dependency on MHD
  - ensemble error decreases with the increasing of MHD only using base learners with 6 hidden units; not clear dependency on MHD in other cases.

ECOC ensemble performance vs. base learner accuracy - 3

Summarizing:

No simple relationships between the ensemble error and the average base learner error with respect to the MHD between codewords

- i.e. if we use fixed length codewords an increment of the MHD does not necessarily lead to improved performances of the ECOC ensemble.

ECOC ensemble performance vs. base learner accuracy – Discussion -1

Explanation: *Different codewords induce different dichotomies*.

The dichotomies can or cannot be hard learnable depending

- on the structure of the data and
- on the type of the base learner used

Complex classification problems require complex dichotomizers ➔ overfitting

The learnability is partially reflected by the average base learner error.

ECOC ensemble performance vs. base learner accuracy – Discussion –2

The ECOC ensemble performance depends on a **complex interaction among:**

- **MHD**
- **accuracy of the dichotomizers**
- **dependency among the codeword bit errors**.

Different experimental cases:

1. the effect due to the error recovering power prevails on the increment of the base learner error
2. the error recovering power is counter-balanced by the increased average base learner error
3. similar trends of the ensemble error and the average base learner error with respect to MHD
4. etc.

Outline

- Output Coding Decomposition Ensembles
- Open problems
- Experimental analysis of factors affecting the effectiveness of Output Coding Decomposition Ensembles
- **Application to Electronic Nose data** (*)
- Conclusions

(*) in collaboration with

Matteo Pardo, *INFM-Brescia (Italy)* and

Giorgio Sberveglieri, *Univ. Brescia (Italy)*

Application to OIL and COFFEE qualification using electronic nose data

Francesco Masulli, *Univ. Pisa (Italy)*
Giorgio Valentini, *Univ. Genoa (Italy)*
Matteo Pardo, *INFM (Italy)*
Giorgio Sberveglieri, *Univ. Brescia (Italy)*

History of Electronic Noses

- **Hartman (1954):** Array of 8 electrochemical cells given different electric patterns to odour samples.
- **Moncrieff (1961):** Array of 6 thermistors covered by different materials, e.g., polymers, gels, and fat. He claimed, even if each thermistor in not selective to odours, the array was able to discriminate a large number of odours
- **Alpha Mos Co. – France** (early '90): Sells the first electronic noses.

Electronic Noses technology to be consolidated

.

Electronic Nose - 1

Task: classification of gas mixtures (such as food flavors, odors)

- A single sensor (e.g., semiconductor thin films)
  - not selective
  - Non-linear and unknown mapping odor-output
- Vectorial data from an array of sensors with different characteristics are more selective

Electronic Nose

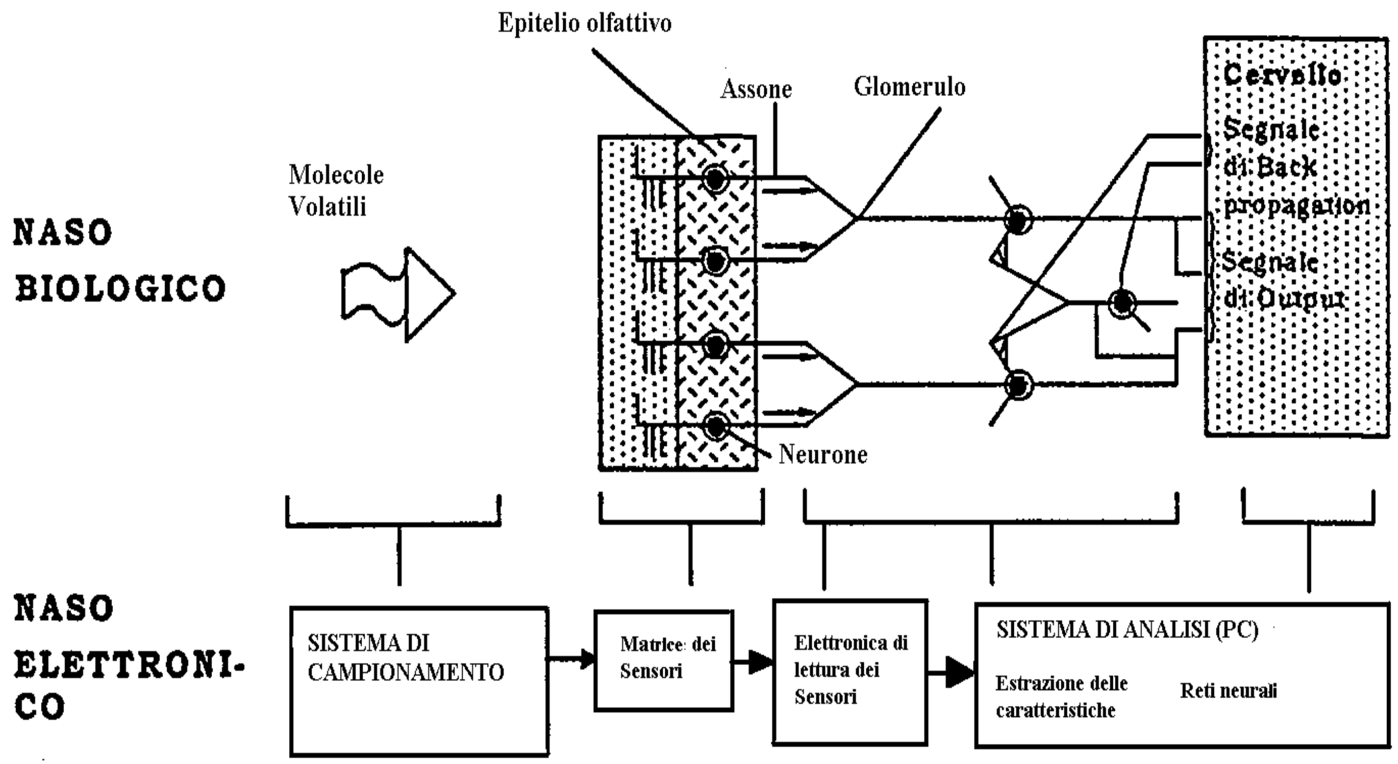
Electronic Noses:

- **array of sensors +**
- **learning machines** (for data analysis)

Data sets characteristics:

- **Few data (100-1000 samples)**
- **Low dimensional (5-10 sensors)**

Human nose ←→ electronic nose

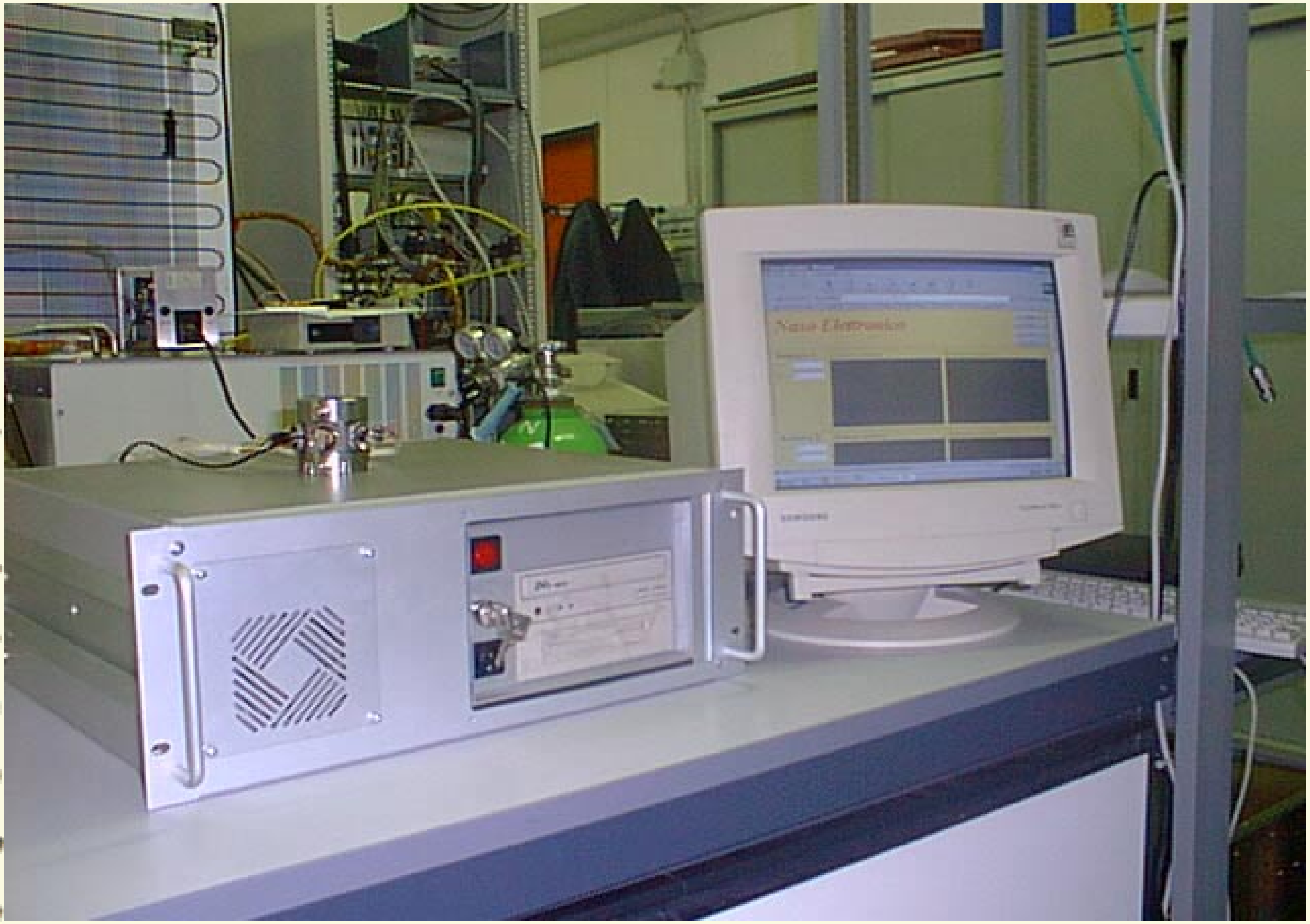


NASO BIOLOGICO

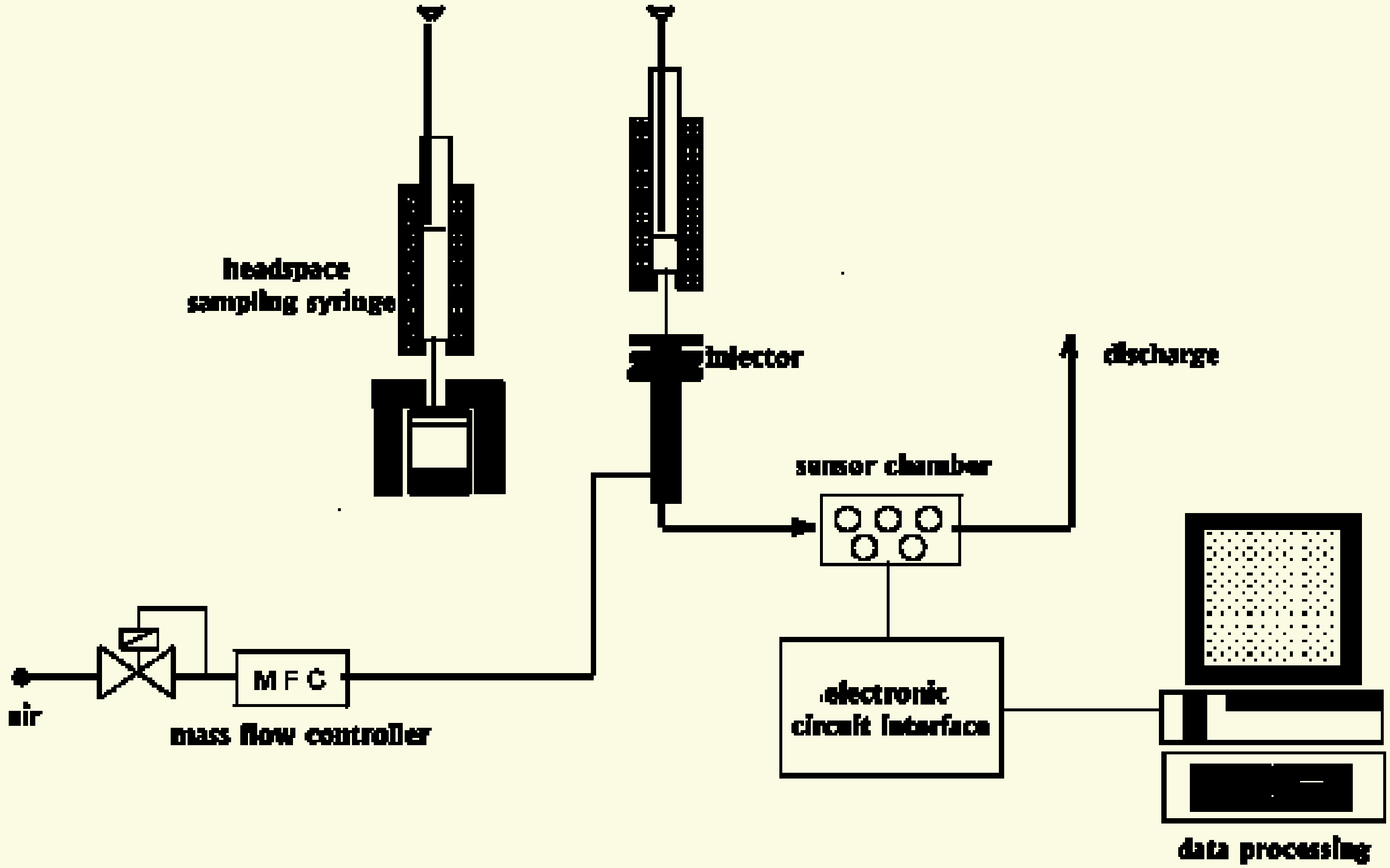
NASO ELETTRONI-CO

SISTEMA DI CAMPIONAMENTO → Matrice dei Sensori → Elettronica di lettura dei Sensori → SISTEMA DI ANALISI (PC): Estrazione delle caratteristiche, Reti neurali

The Pico-1 Electronic Nose



Scheme of PICO1 EN developed at the Gas Sensor Lab - Brescia



Coffee analysis with Pico-1

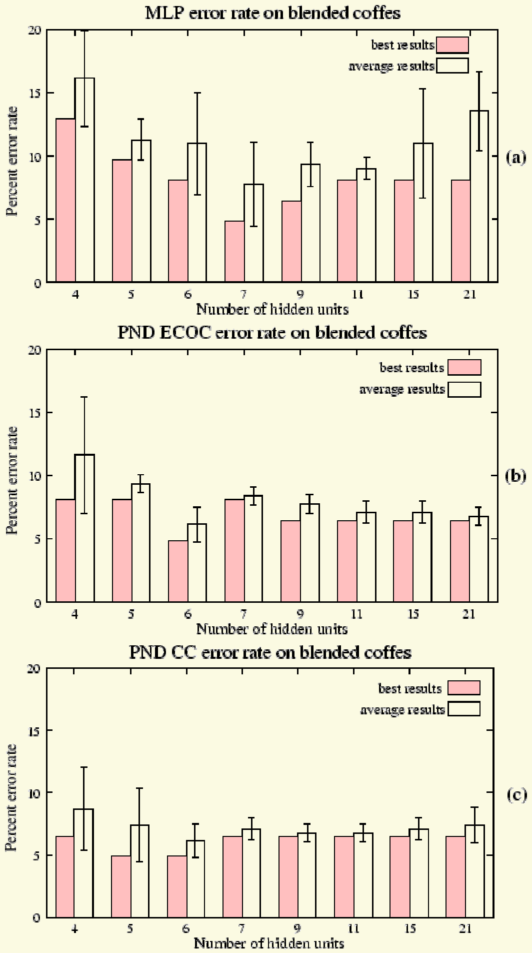
- Two coffee groups:
  - 6 mono varieties + Italian Certified Espresso (ICE)
  - 7 blends for Espresso
- Analysis with Pico-1
- 5 sensors, static sampling, ground coffee, 210 and 249 samples
- Task: mono varieties and blend of coffee discrimination
- Use of PCA + ANN

Blends

249 samples

# coffee	Name	Note, Quality
1	ICE	reference, +
2	ICE, more toasted	strong, +
3	ICE, without natural	study, +
4	Robusta	bad
5	ICE def #1	unripe, -
6	ICE def #2	rancid, -
7	Commercial	arabic + robusta, +-

Blends



Output Coding Decomposition Ensembles show a **lower estimated generalization error** than single learning machines (e.g. Multi-Layer Perceptrons) ➔ accurate systems well-suited to reliable commercial electronic nose devices

Outline

- Output Coding Decomposition Ensembles
- Open problems
- Experimental analysis of factors affecting the effectiveness of Output Coding Decomposition Ensembles
- Application to Electronic Nose data
- **Conclusions**

Conclusions

Output Coding Decomposition Ensembles OPEN PROBLEMS - 1

- Experimental analysis of the trade-off between error recovering capabilities and learnability of the dichotomies induced by the decomposition scheme. Theoretical analyses: Allwein, Shapire & Singer (2000), Valentini (2000).
- Study of the relationship between codeword length and performances. Preliminary results: Ghani (2000).
- Selection of optimal dichotomizers for the DU. Addressed by: Berger (1999), Ghany (2000), Masulli & Valentini (2000).

Output Coding Decomposition Ensembles
OPEN PROBLEMS - 2

- How to design codes jointly maximizing the distance between rows and columns of the DM (a-priori methods).

- How to design codes for a given multiclass problem (a-posterior methods)
  - Greedy approach (Mayoraz & Moreira, 1997)
  - Soft weight sharing (Alpaydin & Mayoraz, 1999)
  - Continuous codes & constrained optimisation problem (Crammer & Singer, 2000)

- How to relate performances of ECOC and dependence among output errors (Kong & Dietterich, 1995; Guruswami & Sahai, 1999).

Output Coding Decomposition Ensembles
OPEN PROBLEMS - 3

We need **more studies to relate the accuracy of the ECOC ensemble with the complexity of the induced decomposition.**

The **relationship effectiveness of ECOC ensemble-complexity of the data** is an item common to other ensemble methods (Ho, 2001) and require specific studies and experimental analysis using appropriate measures of complexity, based on geometrical or topological characteristics of data (Li & Vitanyi, 1993).

end