Análisis del Error en Redes Neuronales: Corrección de los Datos y Distribuciones no Balanceadas

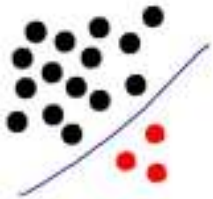
Dr. Roberto Alejo Eleuterio

Índice

- Introducción
- Medidas de evaluación y aspectos experimentales
- Análisis del error: Algoritmo back-propagation con procesamiento por grupos
- Tratamiento del desbalance de las clases con ANN-M
- Corrección de los datos
- Conclusiones, aportaciones y trabajos futuros

Introducción

Problema del desbalance de las clases



Ocurre cuando una o varias clases presentan un número de muestras significativamente mayor respecto al de las otras clases.

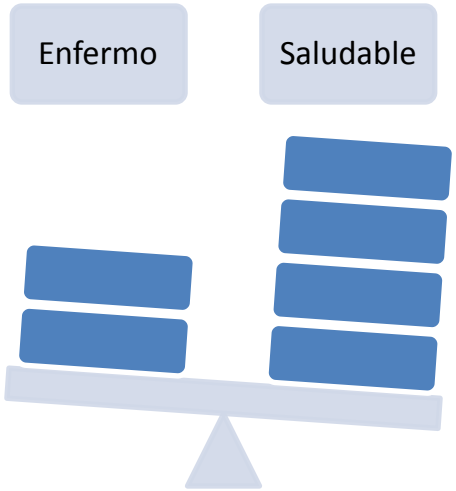
Problemática en redes neuronales

Ocasiona lentitud en la convergencia de las clases minoritarias lo que se traduce en una pobre capacidad de generalización.

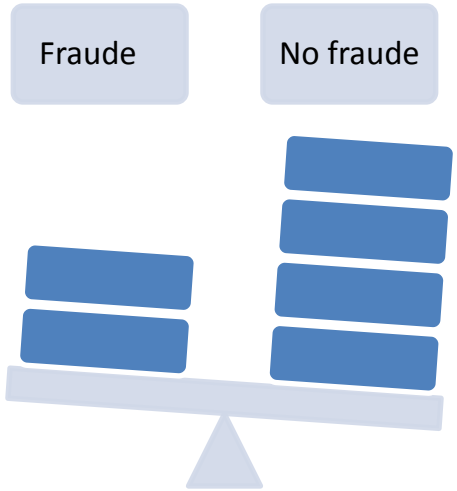
Porqué es importante el estudio del problema?

En general, la clase minoritaria es la que suele contener los casos de interés

Los errores sobre la clase minoritaria suelen suponer un elevado coste

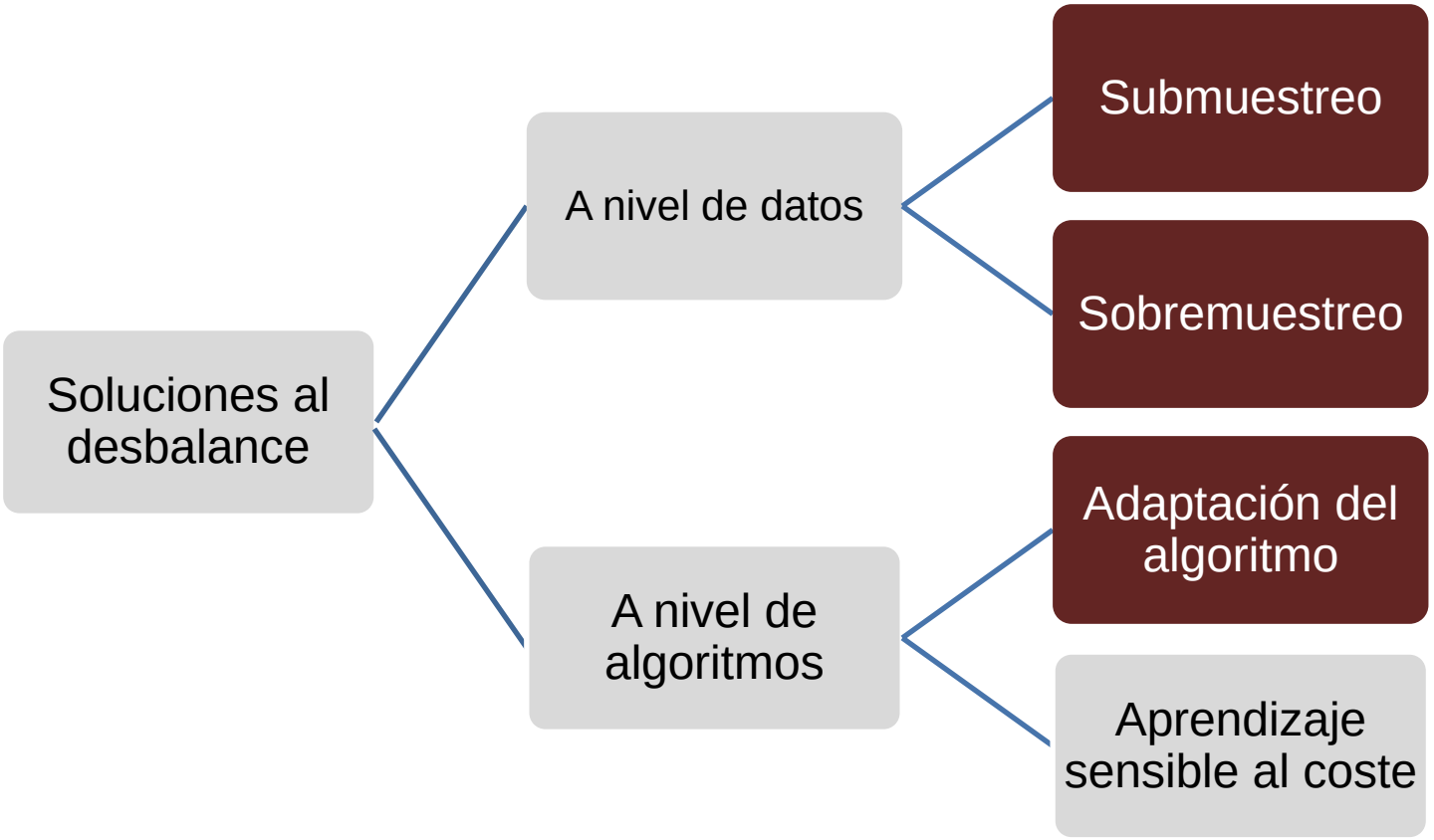


Alto coste en términos de vida



Alto coste monetario

Soluciones al desbalance



Cuán determinante es el desbalance?

Japkowicz y Stephen, 2002

Talla del conjunto de datos + ratio de desbalance

Visa y Ralescu, 2003

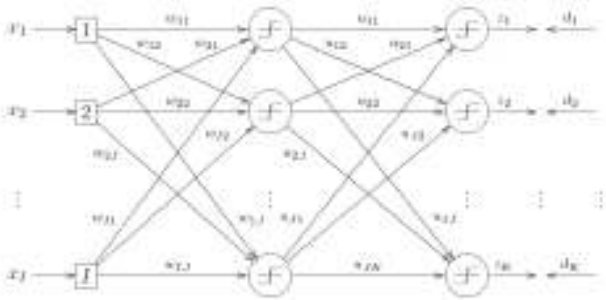
Ratios de desbalance + solapamiento entre clases

Pratti y otros, 2004

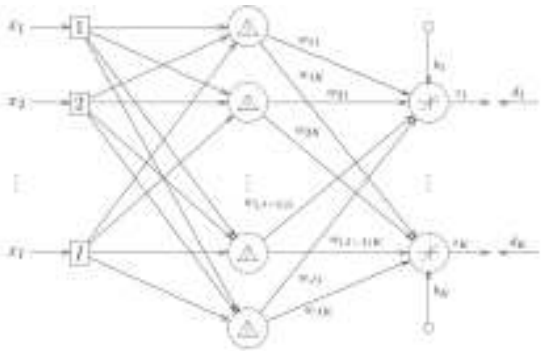
Ratio de desbalance + distancia entre clases

Objetivo

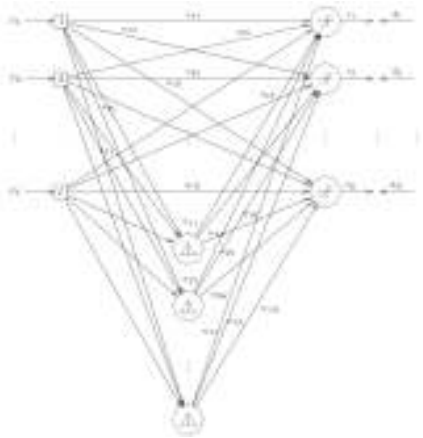
Analizar, estudiar y tratar el problema del desbalance de las clases en el contexto de las redes neuronales artificiales entrenadas con el algoritmo back-propagation con procesamiento por grupos.



MLP



RBF



RBF+VF

Propuestas para tratar el desbalance

1. Inclusión de funciones de coste al proceso de entrenamiento para disminuir los efectos del desbalance de las clases.

2. Descomposición del problema para simplificar el tratamiento del desbalance de las clases a través del uso de redes neuronales modulares.

3. Reducción de la región de solapamiento de las clases minoritarias (mediante técnicas de corrección de los datos) para dar prioridad a las clases minoritarias y así reducir los efectos del desbalance de las clases.

Medidas de evaluación y aspectos experimentales

Métricas de evaluación

Matriz de confusión

	Clases predichas				
Clases reales	1	2	$\cdots$	K	total (N_{i+})
1	N_{11}	N_{12}	$\cdots$	N_{1K}	N_{1+}
2	N_{21}	N_{22}	$\cdots$	N_{2K}	N_{2+}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
K	N_{K1}	N_{K2}	$\cdots$	N_{KK}	N_{K+}
total (N_{+j})	N_{+1}	N_{+2}	$\cdots$	N_{+K}	N

$$PC = \frac{\sum_{i=1}^{K} N_{ii}}{N} \qquad PC_i = \frac{N_{ii}}{N_{i+}} \qquad g-mean = \left(\prod_{i=1}^{K} PC_i\right)^{1/K}$$

Aspectos experimentales

- La experimentación fue realizada en bases de datos de dos y múltiples clases.
- Se aplicó la técnica k-fold-cross-validation.
- La configuración de la red fue determinada según el esquema de prueba y error.
- Cada experimento se ejecutó 10 y 30 veces.
- El criterio de parada de la red fue establecido en 25000 iteraciones o un error inferior a 0.0001.

Análisis del error: Algoritmo back-propagation con procesamiento por grupos

Análisis del error en problemas desbalanceados

Sea K=2 el número de clases, y N el número total de muestras en la ME, entonces el MSE por clase se puede definir como:

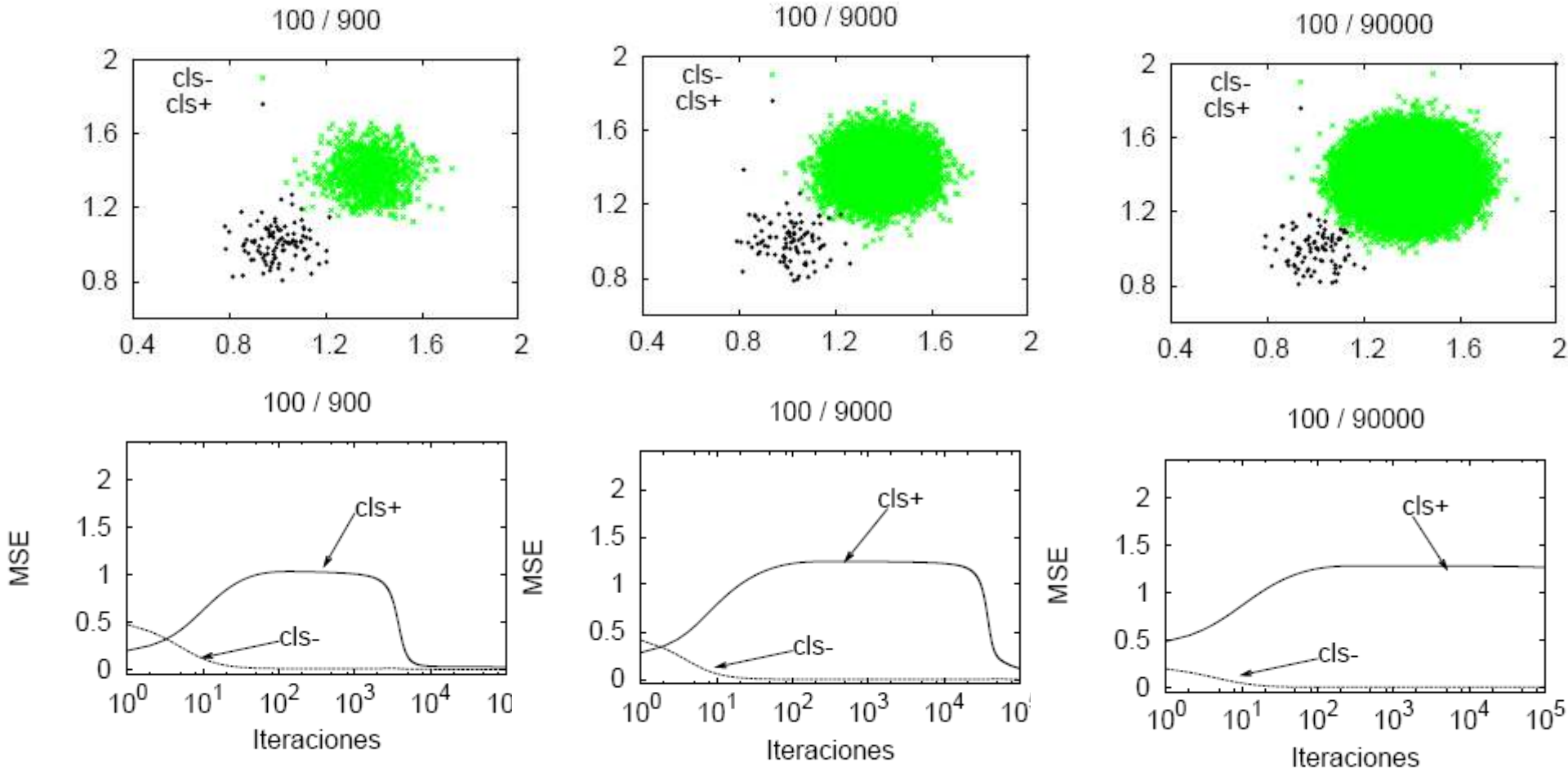
$$E(U)_k = \frac{1}{N}\sum_{n=1}^{N_k}(d^n - f^n)^2,$$ y el MSE global por: $$E(U) = \sum_{k=1}^{K} E(U)_k.$$

Si $N_1 << N_2$, entonces $E(U)_1 << E(U)_2$, y $\|\nabla E(U)_1\| << \|\nabla E(U)_2\|$.

Por lo tanto: $\nabla E(U) \approx \nabla E(U)_2$.

Así, $-\nabla E(U)$ no siempre es la mejor dirección para minimizar el MSE de ambas clases.

Efecto del desbalance de las clases en el MLP



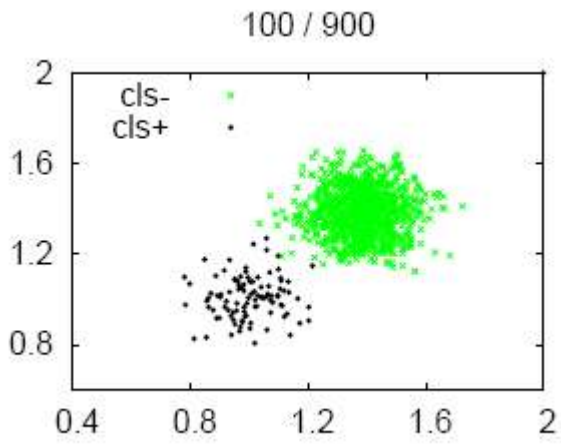
Equilibrio de las aportaciones al MSE

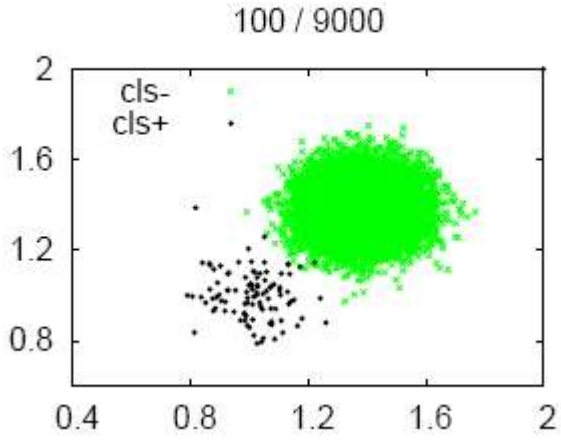
Considérese la inclusión de una función de coste que compense el desequilibrio de las clases como sigue:

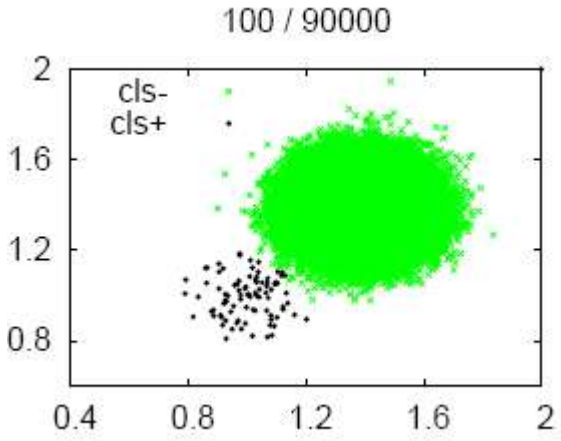
$$\mathrm{E(U)} = \sum_{k=1}^{K} \gamma(k)\mathrm{E(U)}_k = \gamma(1)\mathrm{E(U)}_1 + \gamma(2)\mathrm{E(U)}_2,$$ de esta forma:

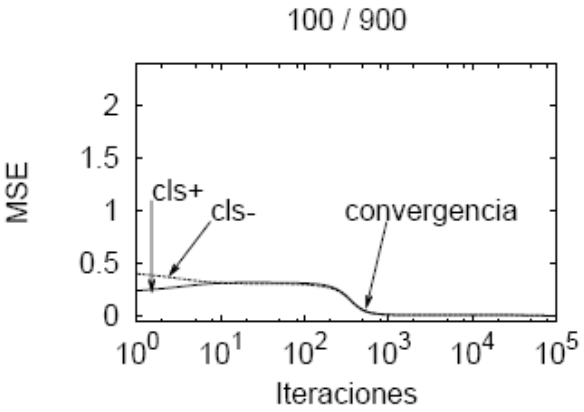
$$\gamma(1)\mathrm{E(U)}_1 \approx \gamma(2)\mathrm{E(U)}_2$$

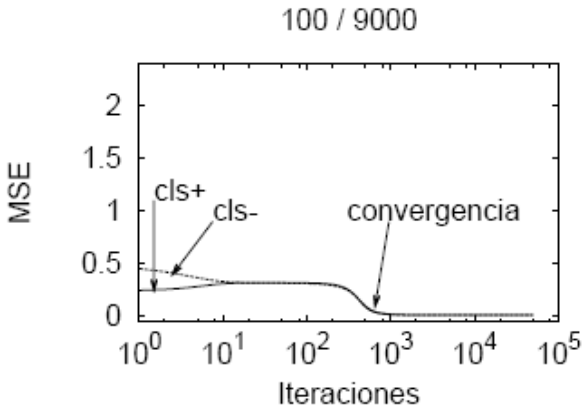
Efecto de equilibrar el error en problemas de dos clases con el MLP

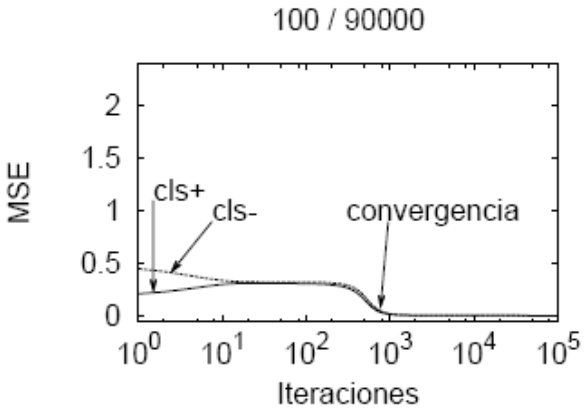












Opciones para tratar el desbalance

- Opción 0: $\gamma(k) = 1$

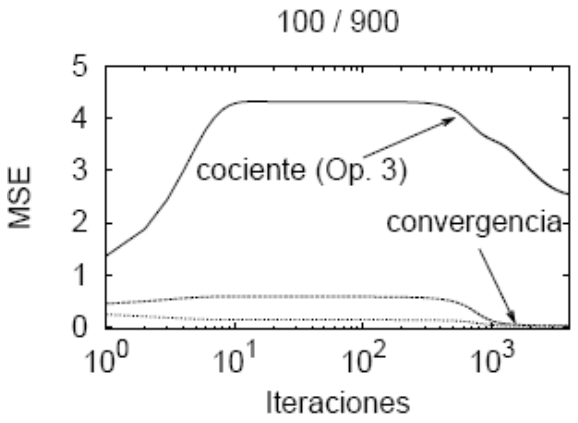
- Opción 1: $\gamma(k) = N_{max} / N_k$
  (Bruzzone 1997)

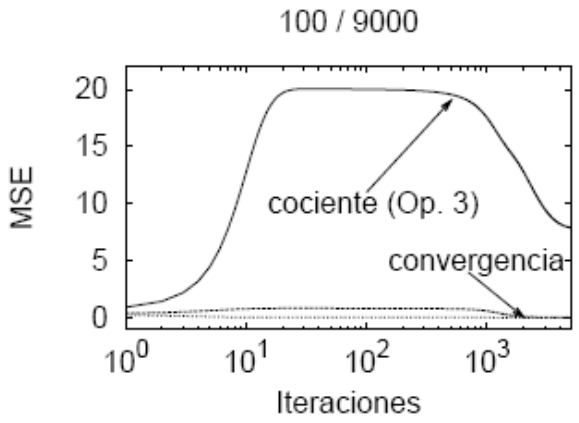
- Opción 2: $\gamma(k) = N_k / N$
  (Fu 2002)

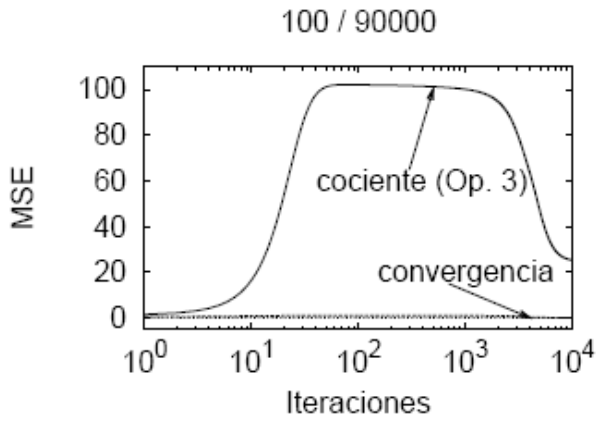
- Opción 3: $\gamma(k) = E(U)_k / E(U)_{max}$
  (Alejo 2008)

Opción 3 con el MLP

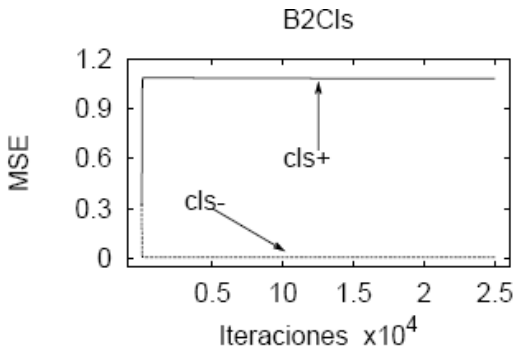
Comportamiento del cociente: $E(U)_k / E(U)_{max}$



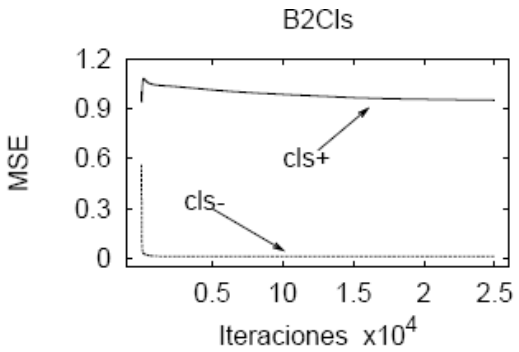




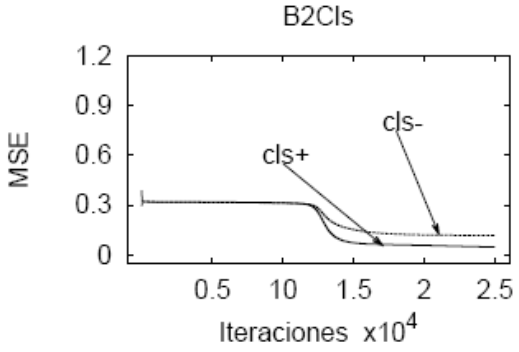
Problemas reales de dos clases



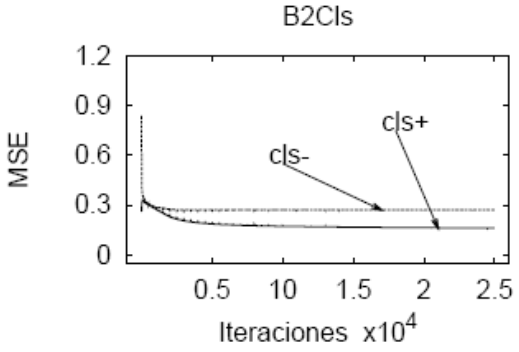
MLP Op. 0



RBF Op. 0



MLP Op. 1

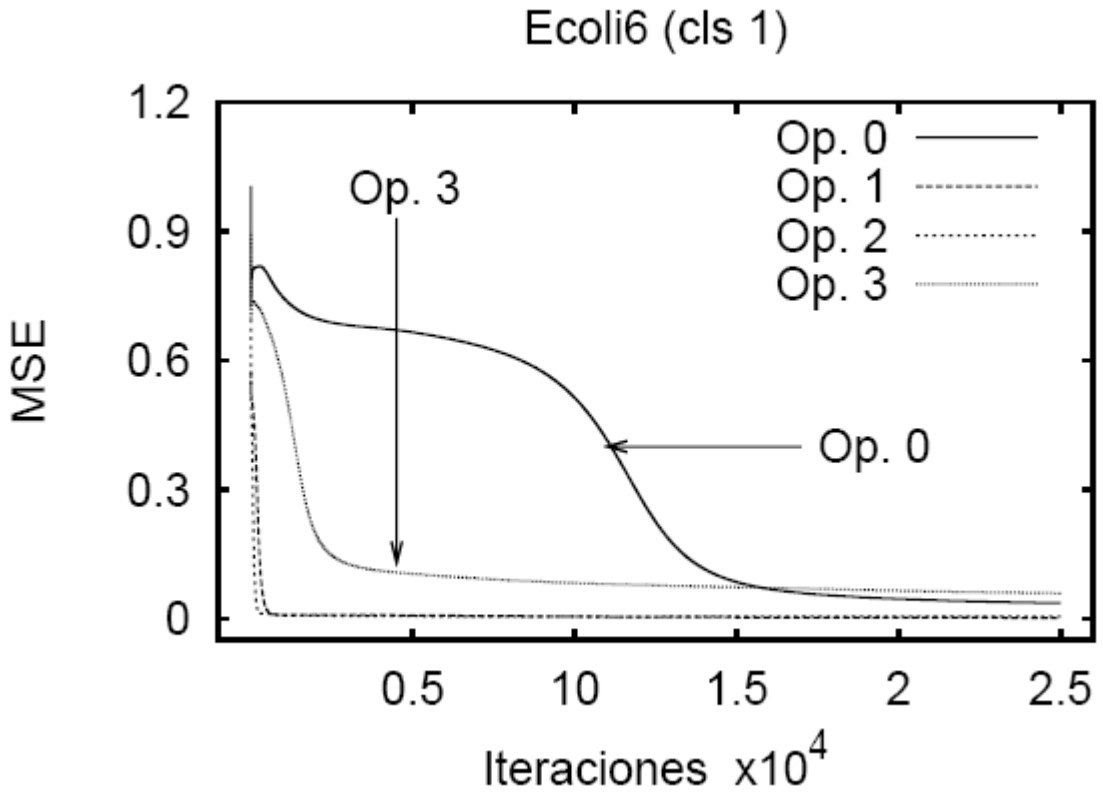


RBF Op. 1

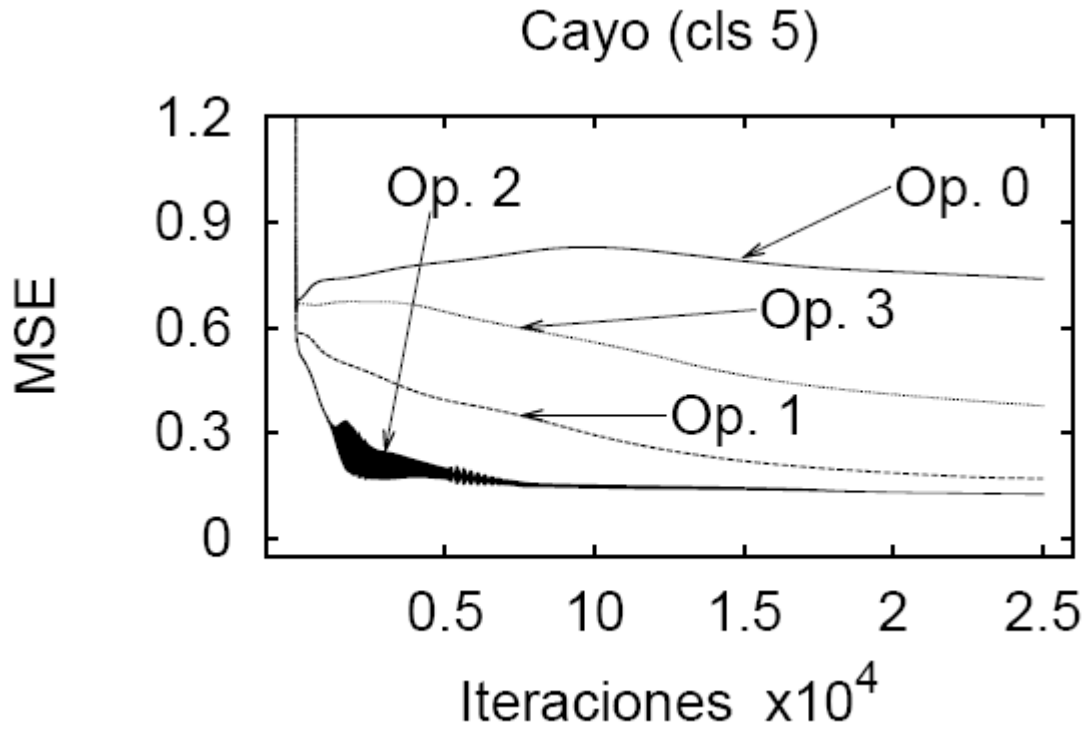
Bases de datos de múltiples clases

- Ecoli6 (poco representada)
- Cayo (suficientemente representada)

Análisis del error en problemas de múltiples clases

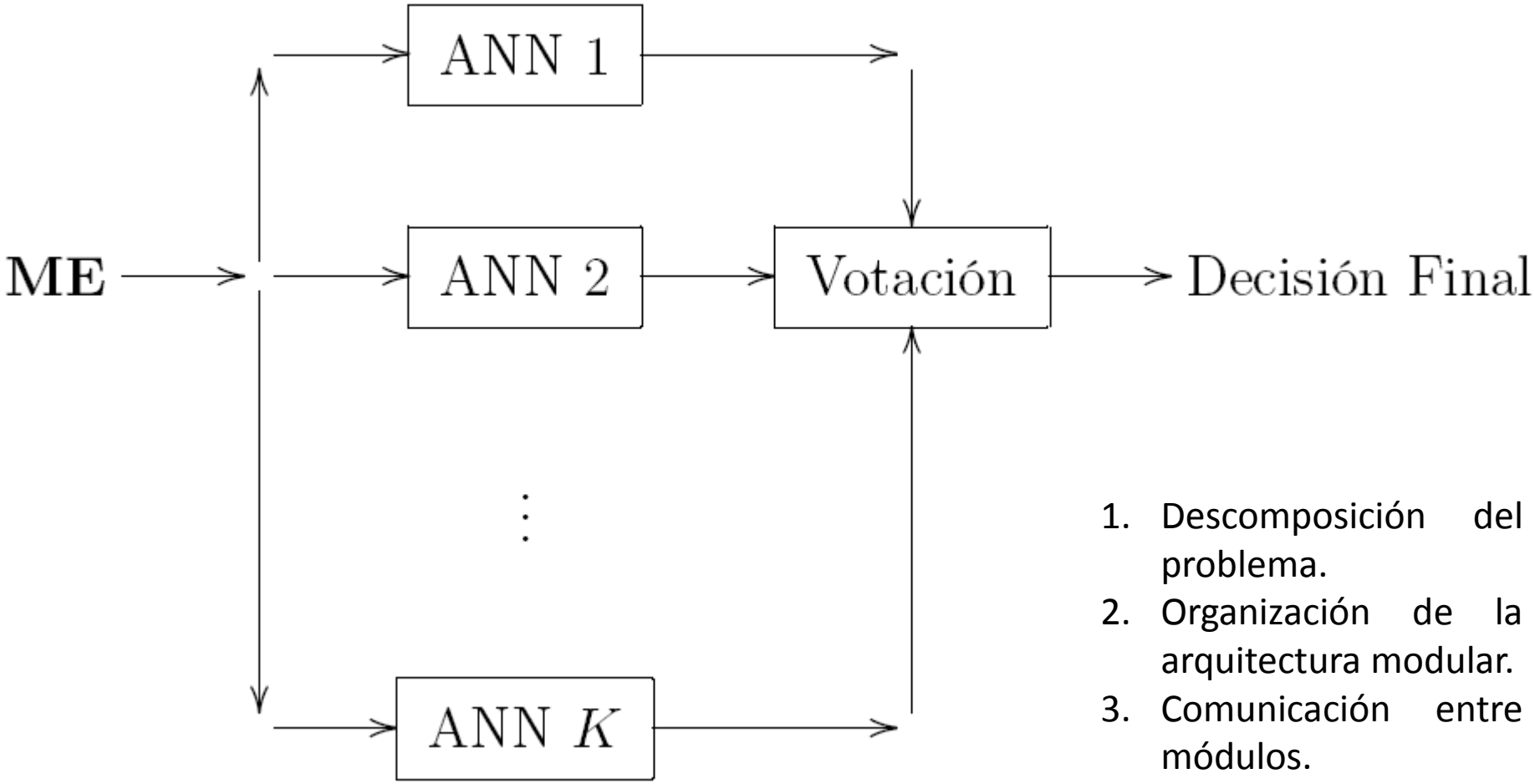


Bases de datos suficientemente representadas



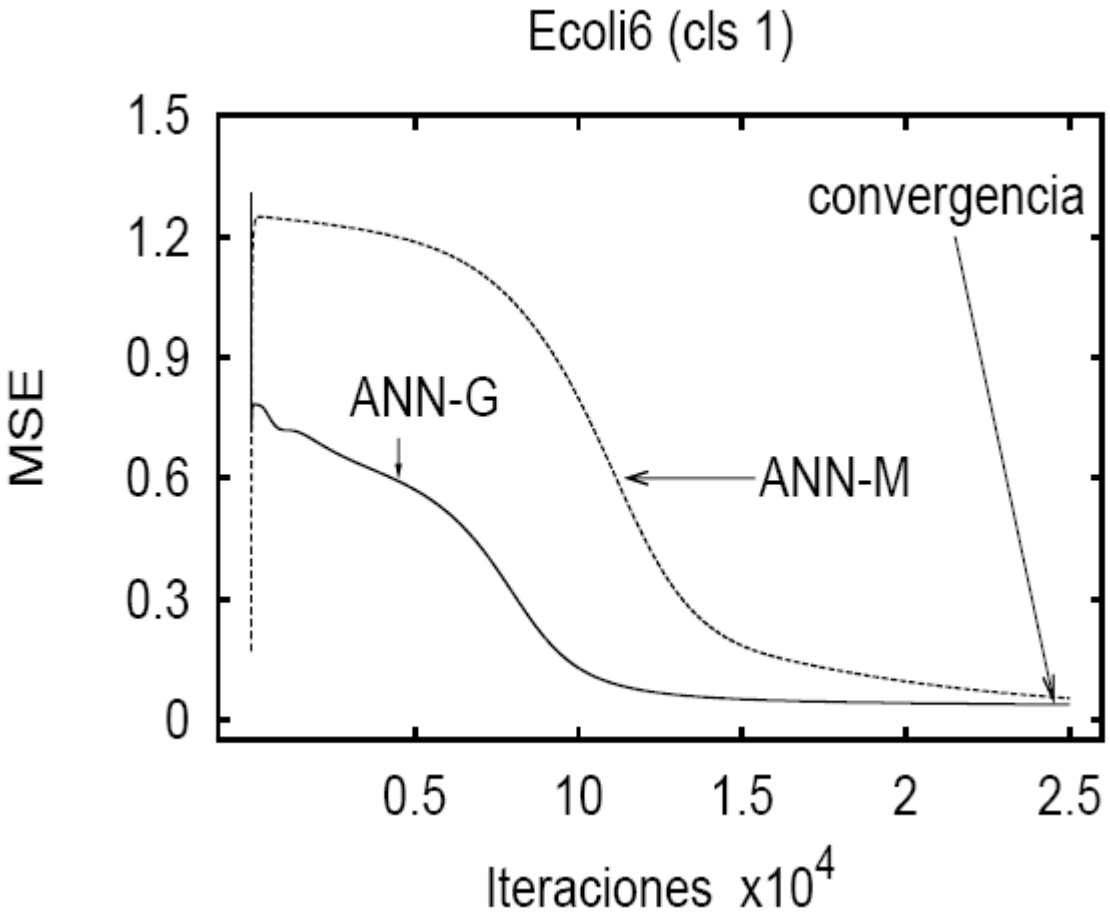
Tratamiento del desbalance de las clases con ANN-M

Esquema simplificado de una red modular

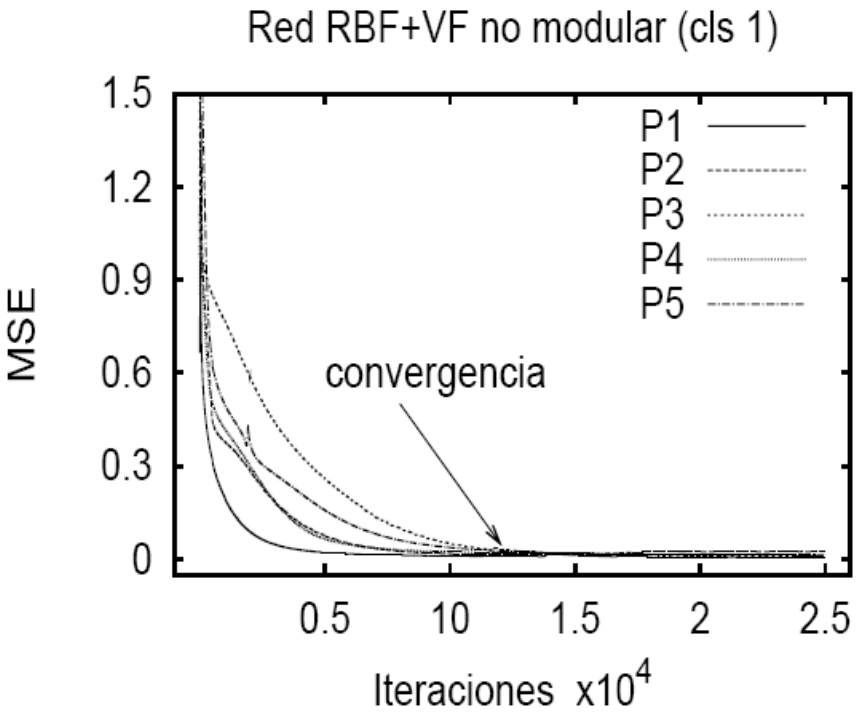


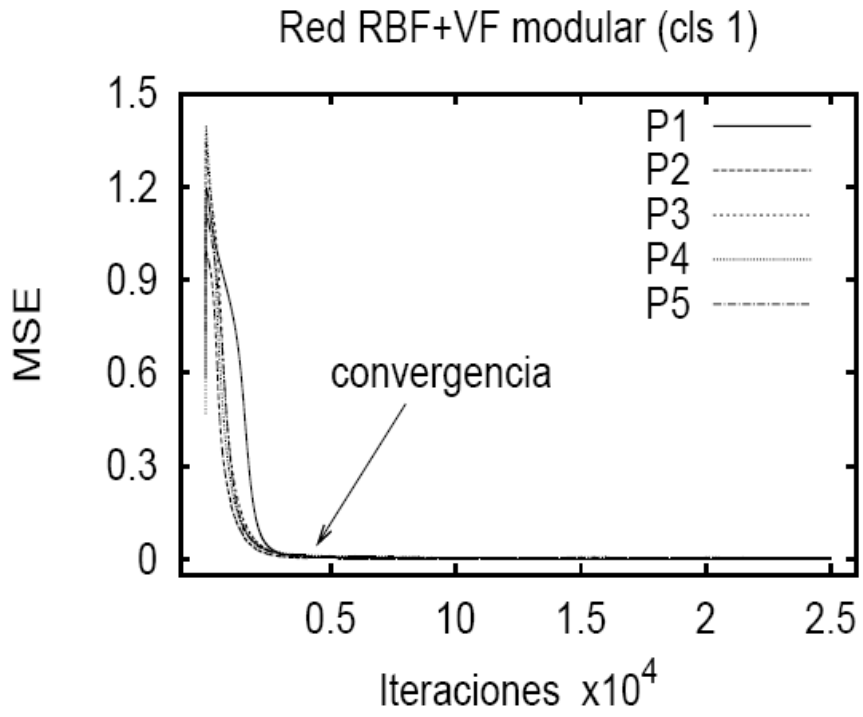
1. Descomposición del problema.
2. Organización de la arquitectura modular.
3. Comunicación entre módulos.

Análisis del error: MLP

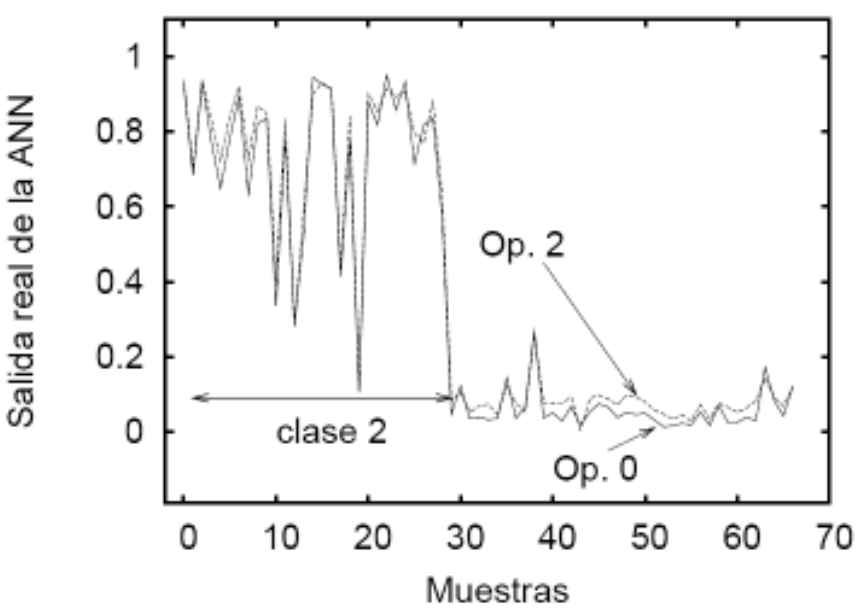


Análisis de error: RBF+VF (Ecoli6)

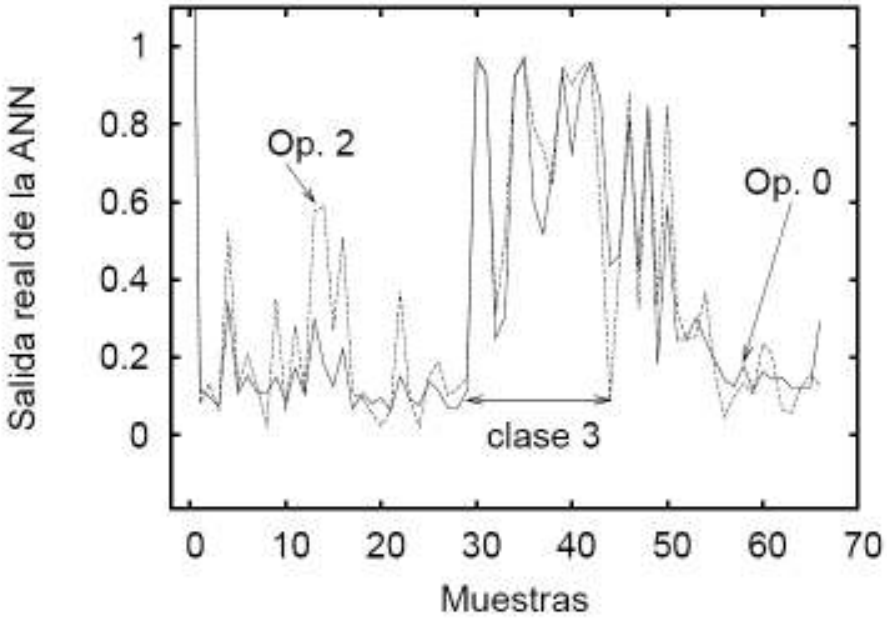




Relación solapamiento - función de coste

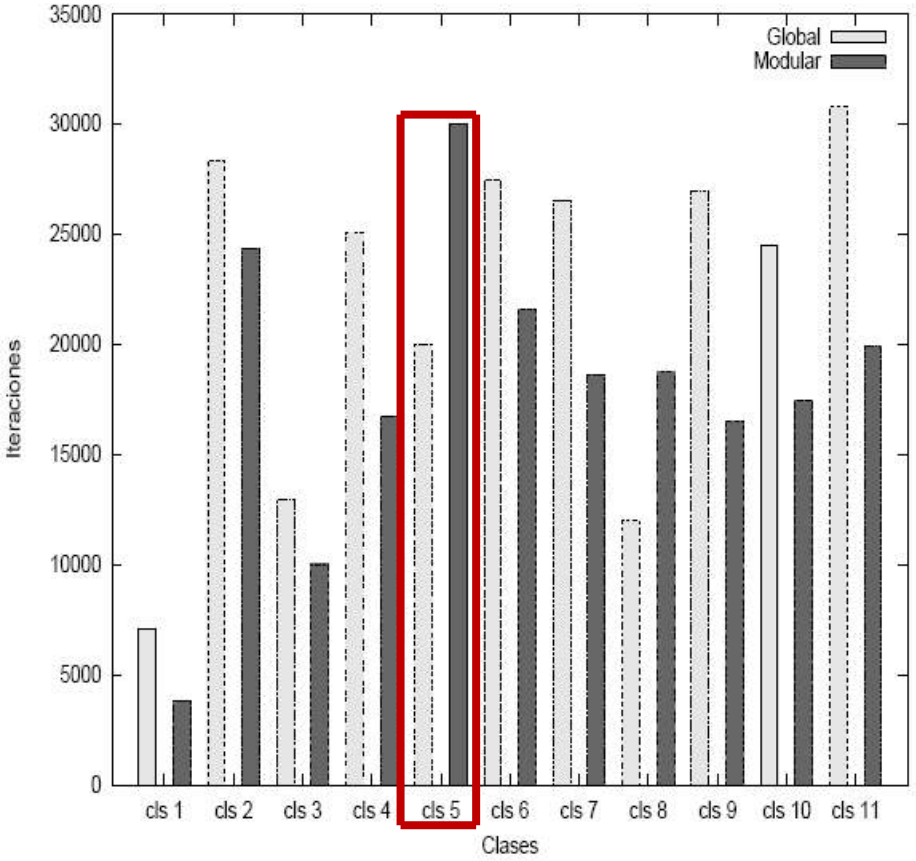


(a) Módulo especializado en la clase 2

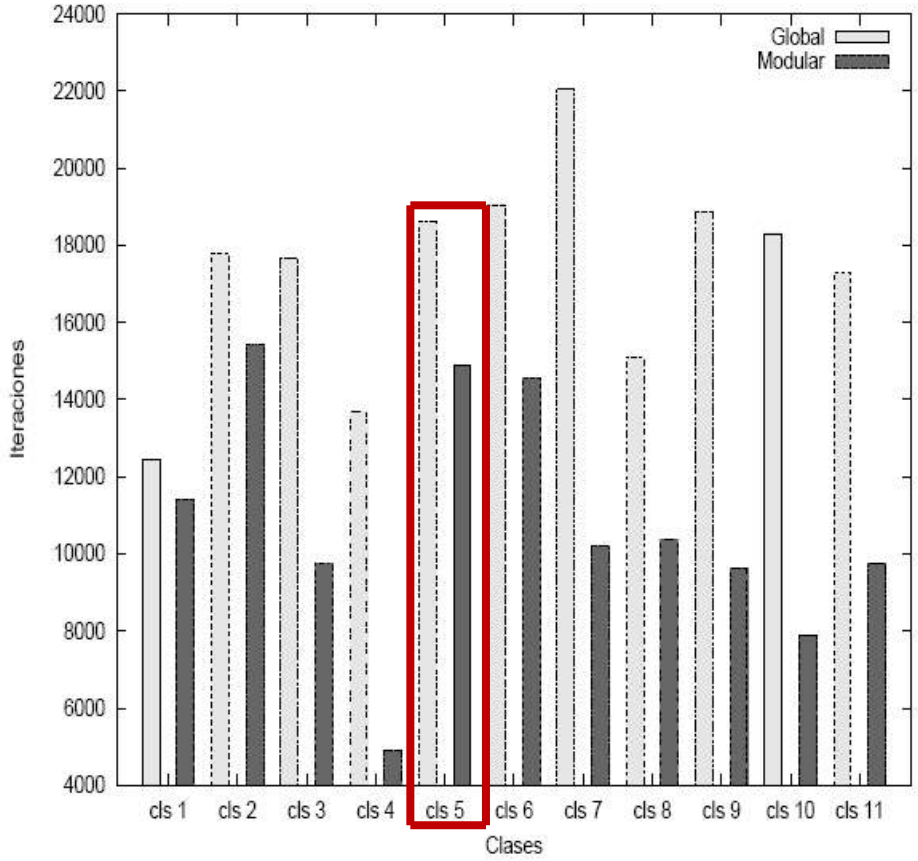


(b) Módulo especializado en la clase 3

Iteraciones para alcanzar un valor mínimo de MSE



(a) ANN-M MLP



(b) ANN-M RBF

Resultados obtenidos con el mecanismo de ponderación

Al ponderar la salida de los distintos expertos se observó que en algunas situaciones se mejora la efectividad del clasificador y ayuda a superar algunos inconvenientes ocasionados por un mal diseño o entrenamiento de los módulos. Sin embargo, no ayuda a mejorar la efectividad en situaciones de solapamiento o falta de representatividad en los datos.

Corrección de los datos

Modificaciones a la edición de Wilson

1. EW $^{-}$

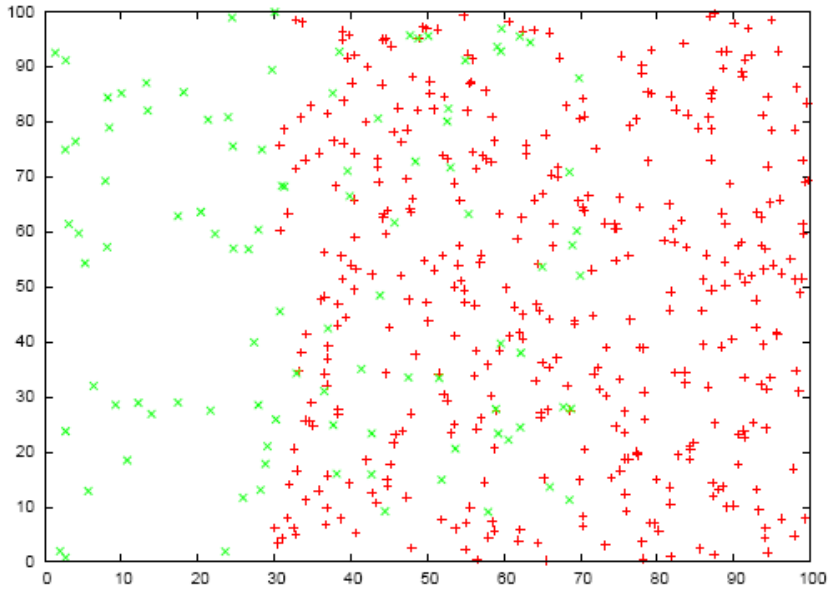
2. EWP

3. EWP $^{-}$

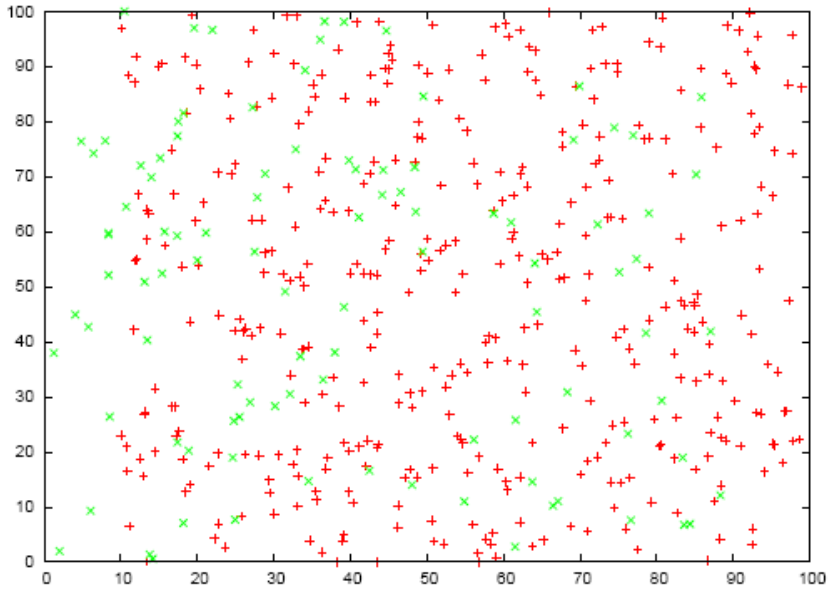
donde la ponderación es establecida como sigue:

$$DP(y,x) = \left(\frac{N_i}{N}\right)^{1/d} \cdot D_E(y,x)$$

Bases de datos sintéticas



(a) 40% de solapamiento



(b) 80% de solapamiento

Resultados dos clases

- Las estrategias que solo eliminan muestras de la clase mayoritaria (EW^{-} y EWP^{-}) mejoran la PC de la clase minoritaria.

- La estrategia EW al eliminar muestras de ambas clases incrementa el desbalance entre clases y reduce la PC de la clase minoritaria.

- La estrategia EWP presenta mejores resultados que EW pero su tendencia es la misma.

Problemas reales de dos clases

	Original		EW		EWP		EW^-		EWP^-	
MLP	PC^-	PC^+	PC^-	PC^+	PC^-	PC^+	PC^-	PC^+	PC^-	PC^+
Diabetes	65.60	81.72	74.80	75.90	63.66	84.89	48.32	90.45	48.14	91.90
German	82.19	55.43	90.70	34.40	86.03	42.23	73.19	66.63	72.00	68.97
Ionosphere	97.56	80.95	98.53	57.30	98.84	63.17	96.93	80.24	96.84	80.63
Phoneme	85.49	62.62	84.90	62.11	83.14	67.01	83.39	68.82	82.13	70.36
RBF	PC^-	PC^+	PC^-	PC^+	PC^-	PC^+	PC^-	PC^+	PC^-	PC^+
Diabetes	60.08	81.64	68.10	77.95	55.70	88.62	34.92	96.79	33.60	96.60
German	92.77	28.03	99.16	4.97	97.66	10.73	83.27	49.80	81.21	52.97
Ionosphere	84.49	92.94	93.20	84.76	90.80	87.86	84.22	93.41	84.31	93.33
Phoneme	88.10	61.38	89.56	56.18	85.66	68.85	85.03	69.81	82.35	75.81
RBF+VF	PC^-	PC^+	PC^-	PC^+	PC^-	PC^+	PC^-	PC^+	PC^-	PC^+
Diabetes	64.54	82.87	75.20	74.18	64.94	84.07	47.90	89.18	44.98	91.46
German	85.76	53.03	92.77	27.97	88.63	38.43	75.59	67.47	74.11	68.03
Ionosphere	95.51	75.87	97.11	61.27	96.93	63.49	93.87	76.98	93.29	75.16
Phoneme	87.79	63.11	88.89	58.98	85.66	68.76	84.39	71.01	82.61	75.62
3-NN	PC^-	PC^+	PC^-	PC^+	PC^-	PC^+	PC^-	PC^+	PC^-	PC^+
Diabetes	77.40	51.87	83.80	48.88	79.60	57.46	64.00	71.64	61.80	72.76
German	82.43	35.00	88.43	26.67	86.71	30.00	70.43	54.00	68.14	56.33
Ionosphere	98.22	59.52	97.78	57.14	97.78	57.14	97.33	66.67	97.33	65.87
Phoneme	93.43	78.25	93.22	73.39	90.70	80.20	89.89	83.10	87.64	85.25

Efectividad del clasificador

Ecoli6

	PC			*g-mean*		
	Op. 0	Op. 0	Op. 3	Op. 0	Op. 0	Op. 3
ANN	Original	EWP$^-$	EWP$^-$	Original	EWP$^-$	EWP$^-$
MLP	86.56(2.78)	86.45(2.43)	84.41(4.42)	83.76(5.57)	83.93(4.58)	83.24(4.41)
RBF	85.87(3.36)	86.64(3.62)	85.13(3.84)	83.50(5.29)	84.43(5.01)	84.18(4.08)
RBF+VF	85.34(3.69)	85.86(3.68)	84.10(4.50)	82.59(5.62)	84.04(4.43)	83.42(4.81)

Cayo

	PC			*g-mean*		
	Op. 0	Op. 0	Op. 3	Op. 0	Op. 0	Op. 3
ANN	Original	EWP$^-$	EWP$^-$	Original	EWP$^-$	EWP$^-$
MLP	83.58(0.77)	84.21(1.10)	85.34(0.41)	70.17(6.28)	71.87(6.88)	81.83(0.65)
RBF	79.9(1.63)	78.12(3.37)	83.69(0.62)	58.97(3.94)	61.90(8.89)	79.53(1.12)
RBF+VF	76.92(2.92)	78.66(4.15)	83.13(0.65)	66.99(7.58)	66.09(8.11)	79.17(0.92)

Conclusiones, aportaciones y trabajos futuros

Distribuciones no balanceadas

1. El desbalance de las clases en la ME impacta en la convergencia de la ANN.

2. El problema viene dado por la desproporción del MSE de las clases minoritarias en relación a las mayoritarias.

Funciones de coste

1. La inclusión de funciones de coste tiene dos consecuencias:
   - Acelerar la convergencia de las clases menos representadas .
   - Reducir la influencia de las clases mayoritarias en el proceso de entrenamiento.

2. Los efectos negativos ocasionados por las funciones de coste son observados principalmente en situaciones donde las bases de datos están poco y/o mal representadas y/o existe alto solapamiento entre clases.

3. La inclusión de funciones de coste al proceso de entrenamiento prioriza a las clases minoritarias lo que se traduce en incrementos de la efectividad del clasificador sobre estas clases.

Tratamiento del desbalance de las clases con ANN-M

1. Descomponer el problema en subproblemas de dos clases reduce la interferencia entre clases, lo que permite que el problema sea más fácil de aprender por la ANN.

2. La descomposición del problema en el caso del MLP puede incrementar el MSE asociado a la clases menos representadas (por la acentuación del desbalance de las clases).

3. El uso de una estrategia adecuada para la combinación de las salidas de los expertos (ANN) puede ayudar a reducir algunas de las deficiencias ocurridas durante la construcción y entrenamiento de los módulos.

4. La tendencia de las ANN-M es producir iguales o mejores resultados de clasificación que los modelos de ANNs globales.

Corrección de los datos

Se observó que reducir la región de confusión a partir de técnicas tomadas del contexto de la regla del vecino más próximo tiene dos efectos:

1. Incrementar la participación de las clases menos representadas en el proceso de entrenamiento.

2. Reducir la influencia de las clases mayoritaria.

Por otro lado, se vio la efectividad de combinar funciones de coste y técnicas de edición.

Aportaciones

1. Presentar un estudio exhaustivo del comportamiento del error en problemas desbalanceados de dos y múltiples clases.

2. Proponer una nueva función de coste para tratar de reducir los efectos del desbalance de las clases.

3. Combinar estrategias distintas para tratar el desbalance de las clases (edición + funciones de coste).

Trabajos futuros

1. Estudiar nuevas funciones de coste que ayuden a acelerar la convergencia de las clases minoritarias y que afecten en menor medida a las clases mayoritarias .

2. Desarrollar nuevos mecanismos que permitan medir el grado de solapamiento entre clases para el contexto de las ANN.

3. Analizar el problema de la falta de representatividad de los datos.

4. Profundizar en el estudio de nuevos mecanismos para lograr un integración efectiva de las salidas de los módulos de una ANN-M.

Análisis del error en redes neuronales: Corrección de los Datos y Distribuciones no Balanceadas

Dr. Roberto Alejo Eleuterio