

## Análisis del error en redes neuronales artificiales: Distribuciones no balanceadas

R. Alejo<sup>1,2,3</sup>, P. Toribio<sup>2</sup>, F. García<sup>2</sup>

<sup>1</sup> Dept. Llenguatges i Sistemes Informàtics, Universitat Jaume I  
Av. Sos Baynat s/n, 12071 Castelló de la Plana (España)

<sup>2</sup> Centro Universitario Atlacomulco, Universidad Autónoma del Estado de México  
Carretera Toluca-Atlacomulco Km. 60 Atlacomulco (México)

<sup>3</sup> Lab. de Reconocimiento de Patrones, Instituto Tecnológico de Toluca  
Av. Tecnológico s/n, 52140 Metepec (México)  
Paper received on 04/08/08, accepted on 06/09/08.

**Resumen.** En este trabajo se presenta un estudio empírico comparativo de diferentes estrategias diseñadas para reducir la influencia del desequilibrio de las clases en los datos de entrenamiento. Se centra en el aprendizaje y clasificación de las redes neuronales artificiales perceptrón multicapa y de funciones de base radial. Estas redes son entrenadas con el algoritmo back-propagation con procesamiento por grupos. El estudio se realizó sobre bases de datos de dos clases en tres escenarios distintos.

### 1 Introducción

La entrada a una Red Neuronal Artificial (RNA) con aprendizaje supervisado, consiste de una Muestra de Entrenamiento (ME). Una ME, es un conjunto de datos previamente identificados por un experto humano [1], que caracteriza un problema a resolver.

Generalmente, los métodos de aprendizaje supervisado, como las RNA, están diseñados para trabajar con MEs razonablemente balanceadas [2], es decir, donde la diferencia en el número de patrones de las distintas clases no es considerable. Sin embargo, existen numerosas aplicaciones donde la desproporción en el número de patrones entre clases es significativa [3]. Por ejemplo, en la detección de fraudes en llamadas telefónicas [4], en la identificación de productos defectuosos en la línea de ensamblaje de partes de automóviles [5], o en transacciones con tarjetas de crédito [6] (el número de transacciones legales es mucho mayor que el número de transacciones ilegales). También se han presentado en problemas médicos o en el diagnóstico de enfermedades raras.

Actualmente, el problema del desequilibrio de las clases de la ME, ha sido considerado un problema crítico en la minería de datos y el aprendizaje automático [7].

En el contexto de las RNA del tipo Perceptrón Múlticapa (PM), entrenadas con el algoritmo back-propagation y dominios de dos clases el problema se ha formulado como sigue: la clase mayoritaria domina el proceso de entrenamiento, y los elementos de la clase menos representada o minoritaria pueden ser ignorados, de forma que la convergencia de la clase minoritaria es muy lenta [8].

Diversas propuestas se han encaminado a tratar este problema. En [9], se analiza al algoritmo back-propagation y se propone su modificación para acelerar el proceso de convergencia de la red. La idea está centrada en el cálculo del vector gradiente y su dirección, de tal forma, que permita que el error pueda decrecer en dirección de ambas clases y se evite que la clase minoritaria sea ignorada.

En [8], se expone una idea similar para modificar el algoritmo back-propagation y acelerar su convergencia. Esta modificación consiste básicamente en incluir una función de costo en el algoritmo de entrenamiento y disminuir su valor a partir de una estrategia heurística, de tal forma que se reduzca su impacto en la probabilidad de la distribución de los datos. Sin embargo, las estrategias más estudiadas para enfrentar el problema del desbalance en la ME, han sido las técnicas de sub-muestreo (*under-sampling*) o sobre-muestreo (*over-sampling*) [1].

En este trabajo, se estudian tres estrategias diseñadas para equilibrar el desbalance de la ME durante el proceso de entrenamiento de los modelos neuronales RNA-PM, y RNA-FBR. El estudio se realizó en dominios de dos clases con RNA entrenadas con el algoritmo back-propagation en *batch mode* o procesamiento por grupos.

## 2 Análisis del error

Estudios empíricos realizados al algoritmo back-propagation [9], muestran que el desbalance de las clases de la ME genera aportaciones desiguales al error cuadrático medio (mean square error, MSE) en la fase de entrenamiento de la RNA, es decir, la mayor parte de las aportaciones al MSE están dadas por la clase mayoritaria. En consecuencia, el entrenamiento de la red es dominado por esta última.

Consideremos una ME de dos clases ( $m = 2$ ) con  $N$  patrones de entrenamiento.

$N = \sum_i^m n_i$ , tal que  $n_i$  sea el número de patrones de la clase  $i$ . Supongamos que el MSE por clase puede ser expresado como

$$E_i(U) = \frac{1}{N} \sum_{n=1}^m \sum_{p=1}^L (y_p^n - F_p^n)^2, \quad (1)$$

de tal forma el MSE global puede ser expresado como

$$E(U) = \sum_{i=1}^m E_i(U) = E_1(U) + E_2(U). \quad (2)$$

Si  $n_1 \ll n_2$  entonces  $E_1(U) \ll E_2(U)$  y  $\|\nabla E_1(U)\| \ll \|\nabla E_2(U)\|$ . Por lo tanto  $\nabla E(U) \approx \nabla E_2(U)$ . Así,  $-\nabla E(U)$  no siempre es la mejor dirección para minimizar el MSE de ambas clases [9].

Considerando que el problema del desbalance de la ME afecta negativamente al algoritmo back-propagation, debido a la desproporción de las aportaciones al MSE (Ec. 2) por parte de las clases, se puede considerar la inclusión de una función de costo ( $\Psi$ ) al algoritmo, que compense el desequilibrio de las clases de la ME como se muestra a continuación:

$$E(U) = \sum_{i=1}^m \Psi(i) E(U)_i = \Psi(1) E_1(U) + \Psi(2) E_2(U) . \quad (3)$$

de esta forma  $\Psi(1) \parallel \nabla E_1(U) \parallel \approx \Psi(2) \parallel \nabla E_2(U) \parallel$  y puede evitarse que la clase minoritaria sea ignorada y el entrenamiento sea dominado por la clase mayoritaria. En este trabajo se estudian las siguientes funciones de costo:

Opción 0:  $\Psi(i) = 1$ , es decir, el algoritmo back-propagation sin ninguna modificación.

Opción 1:  $\Psi(i) = n_{\max} / n_i$ ; donde  $i = 1, \dots, m$ ; es el total de clases y  $n_{\max}$  es el número de patrones de la clase mayoritaria.

Opción 2:  $\Psi(i) = N / n_i$ ; donde  $i = 1, \dots, m$  y  $N$  es el número total de patrones.

Opción 3:  $\Psi(m) = \parallel \nabla E_{\max}(U) \parallel / \parallel \nabla E_m(U) \parallel$ , donde  $\parallel \nabla E_{\max}(U) \parallel$  corresponde a la clase mayoritaria. Esta función es una simplificación de la propuesta de [9].

### 3 Estudio del MSE durante el entrenamiento de la red

Para evaluar las posibilidades de las estrategias presentadas en la sección 2, se realizó un estudio empírico comparativo del comportamiento del MSE por clase en tres situaciones diferentes. Se utilizaron las bases de datos identificadas como: V2Cls, Phoneme y B2Cls.

V2Cls y B2Cls corresponden a las bases de datos Vowel y Balance respectivamente. Cada conjunto de datos se transformó a conjuntos de dos clases. En cada una de ellas se tomó la clase 1 como minoritaria y el resto de las clases formaron la clase mayoritaria. De esta forma se obtuvieron bases de datos desbalanceadas de dos clases. La finalidad fue la de generar problemas prototípico no balanceados que representaran los diferentes escenarios que ocurren comúnmente. El uso de bases de datos de dos clases permite simplificar la interpretación del MSE por clase.

Tabla 1. Características de las bases de datos

Datos	Tamaño	Atributos	Distribución	Ratio
V2Cls.	528	10	48/480	0.100
Phoneme	5404	5	1586/381	0.415
B2CLS	625	4	49/576	0.085

En la Tabla 1 se presentan las características más relevantes para esta investigación de los conjuntos de datos V2Cls, Phoneme y B2Cls. Se observa, que las tres bases datos muestran niveles diferentes de desequilibrio en las clases, además de presentar tres situaciones de ocurrencia común.

La evaluación de las estrategias Opción 1, 2 y 3 se realizó en el contexto de las RNA-PM y RNA-FBR entrenadas con el algoritmo back-propagation con procesa-

miento por grupos. Los parámetros libres de la red fueron identificados a partir de estrategias de prueba y error.

En las Fig. 2, 3 y 6, se presenta el MSE de ambas clases (MSE+, para la clase positiva o minoritaria y MSE-, para la clase negativa o mayoritaria) para los tres modelos de RNA y funciones de costo (Opciones 0-3). El eje x representa las iteraciones, mientras que el eje y corresponde al MSE.

Se observa en la Fig. 2a y 2b el comportamiento del MSE por clase cuando la ME esta desbalanceada y el algoritmo de entrenamiento no es modificado.

Inicialmente el MSE de la clase de la clase mayoritaria decrece rápidamente (en las primeras iteraciones), mientras que el MSE de la clase minoritaria es incrementado. Posteriormente, el MSE de esta clase decrece muy lentamente.

Es interesante notar que la RNA-PM alcanza la convergencia mucho antes que la RNA-FBR, en el caso de esta última no se logra la convergencia y se ve afectada en mayor medida por el desequilibrio de las clases como se ilustra en la Fig. 2b.

Al incluirse las Opciones 1, 2 y 3 al proceso de entrenamiento (ver Fig. 2c-h) se observa que el MSE de ambas clases decrece uniformemente, evitando que el entrenamiento sea dominado por la clase más representada en la ME.

El efecto de las Opciones en la convergencia de la red es notable. Se muestra en la Fig. 2 un considerable incremento en la velocidad de convergencia de las RNA, en especial en el modelo PM.

Por otra parte, en términos de clasificación (Acc) y valores de g-mean, se observan porcentajes cercanos al 100% tanto con la Opción 0 como con el resto de las Opciones (ver Fig. 2), es decir, estos valores no fueron incrementados ni disminuidos (significativamente) al aplicar las Opciones 1, 2 y 3, se incrementó la velocidad de convergencia de la clase minoritaria.

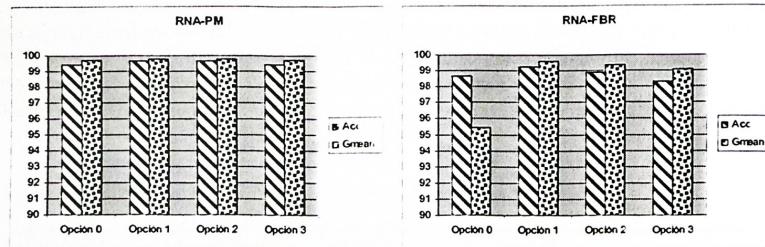


Fig. 1. En la figura se esquematiza el desempeño del clasificador con los datos V2C1s.

En la Fig. 3, se presenta el comportamiento del MSE por clase de la base de datos Phoneme. Se observa una conducta diferente a la presentada por la base de datos V2C1s. Puede verse en los incisos (a) y (b), que el MSE de la clase minoritaria es mucho mayor que el de la mayoritaria durante el proceso de entrenamiento y además no se alcanza la convergencia de la RNA en ninguno de los modelos.

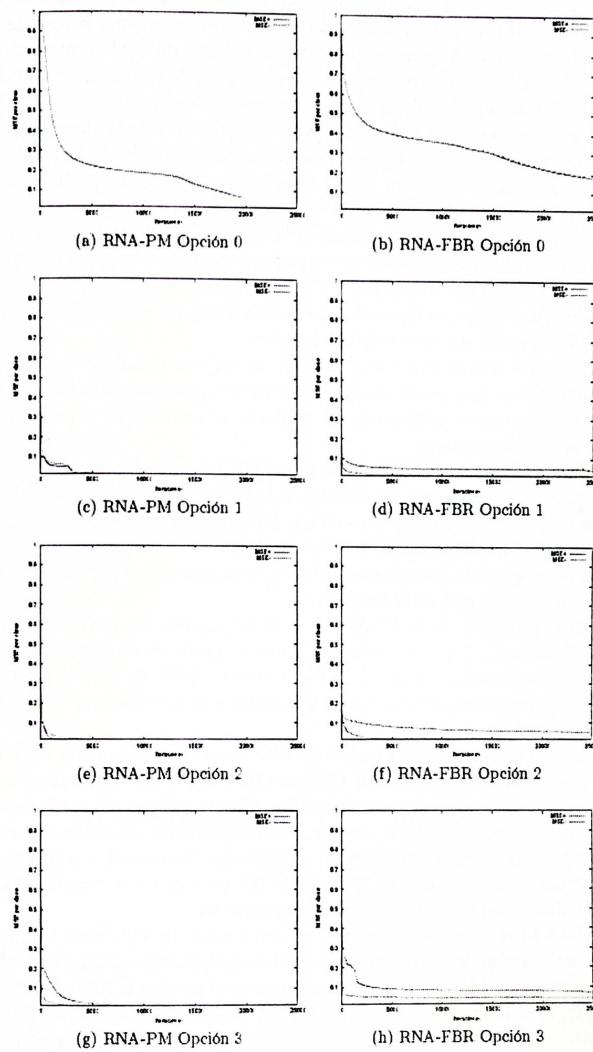


Fig. 2. En la figura se ilustra el MSE por clase de la base de datos V2Cl.

En la Fig. 3, se presenta el comportamiento del MSE por clase de la base de datos Phoneme. Se observa una conducta diferente a la presentada por la base de datos V2Cls. Puede verse en los incisos (a) y (b), que el MSE de la clase minoritaria es mucho mayor que el de la mayoritaria durante el proceso de entrenamiento y además no se alcanza la convergencia de la RNA en ninguno de los modelos.

Al aplicarse las Opciones 1, 2 y 3, el porcentaje de error de la clase minoritaria es disminuido considerablemente, no obstante, la convergencia de la RNA no se consigue. Es interesante notar que no se observa una diferencia significativa entre los resultados arrojados por la aplicación de las Opciones 1, 2 y 3.

Por otra parte, al evaluar el desempeño del clasificador a partir de los criterios de medida: precisión (Acc) y, valores de  $g$ -mean se percibieron resultados importantes. En la Fig. 4, se aprecia que con la RNA-PM en términos de precisión no hubo diferencia estadística (excepto en Opción 3), lo que significa que no fue afectada la precisión de la RNA por la aplicación de las Opciones.

En cuanto a los valores de  $g$ -mean se observan mejoras estadísticamente significativas. Valores altos de  $g$ -mean implica un buen desempeño del clasificador sobre ambas clases (minoritaria y mayoritaria), es decir, se incremento el porcentaje de aciertos de la clase minoritaria.

La Fig. 6 correspondiente a la base de datos B2Cls muestra una visión diferente a las presentadas anteriormente. Se observa en la Fig. 6a y 6b, que el MSE de clase minoritaria se incrementó en mayor medida y decreció más lentamente en comparación con las bases de datos anteriores. Esto trajo como consecuencia que la clase minoritaria fuera ignorada durante el proceso de entrenamiento y por lo tanto el porcentaje de aciertos para esta clase fuera cero.

En el caso particular de la RNA-PM se observan mejoras muy significativas cuando las Opciones 1, 2 y 3 son aplicadas durante el proceso de entrenamiento. Se observa en los incisos (c), (e) y (g) de la Fig. 6 como el MSE de la clase minoritaria es reducido de manera importante, lo que trae como resultado mejoras en el desempeño del clasificador.

En términos de precisión (Acc) no existe diferencia estadística entre los resultados obtenidos por la RNA-PM con las Opciones 0, 1, 2 y 3, mientras que en valores de  $g$ -mean se tienen incrementos considerables. Obsérvese que con la Opción 0 se tienen valores de  $g$ -mean iguales a cero, i.e., no es identificada correctamente en ninguno de los casos (Fig. 5). No obstante, al aplicarse las Opciones 1, 2 y 3 el porcentaje de aciertos para esta clase se incrementa (en promedio) al menos 90%. Así mismo la confiabilidad de la clasificación es incrementada.

En la RNA-FBR, al igual que ocurrió en otros casos las Opciones 1, 2 y 3 incrementan la velocidad de convergencia de la clase menos representada, no obstante, no se tienen resultados de similar importancia que en el caso de la RNA-PM.

En la Fig. 5, se observa que con la RNA-FBR el porcentaje de aciertos de la clase minoritaria es incrementado al aplicarse las Opciones 1 y 2, pero también es notable la reducción de la precisión. Por otro lado, la Opción 3 no parece ayudar a incrementar el rendimiento del clasificador.

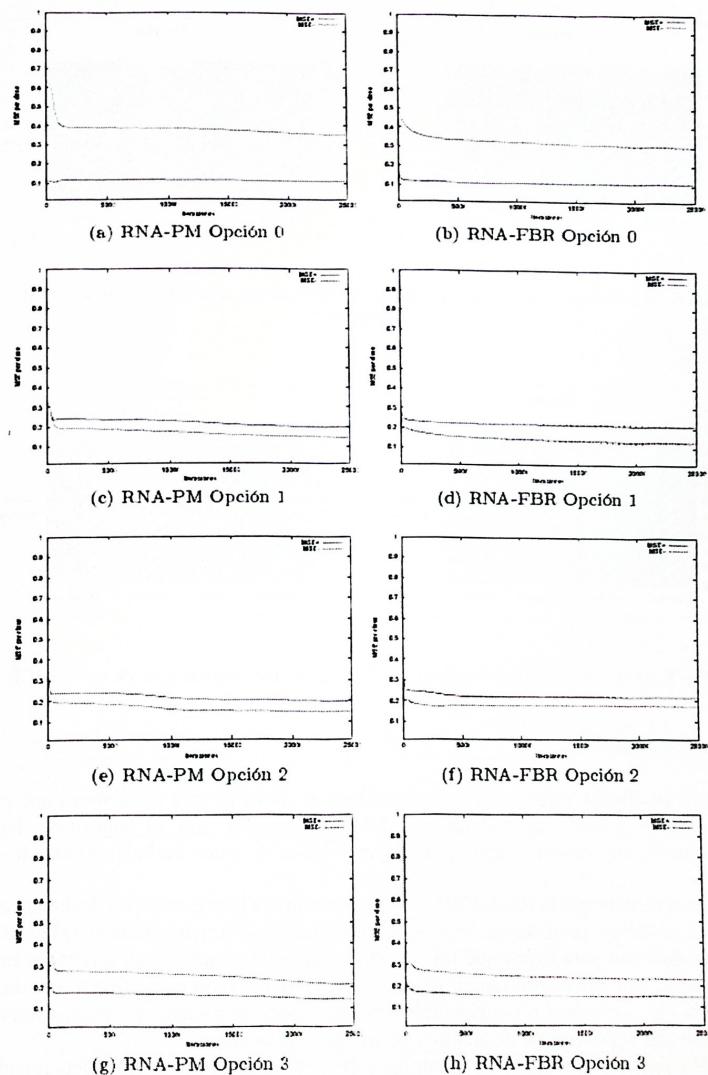


Fig. 3. En la figura se ilustra el MSE por clase de la base de datos Phoneme.

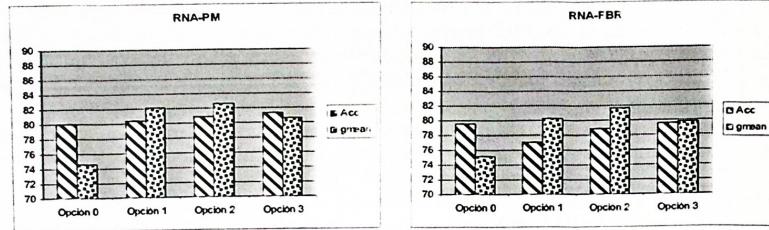


Fig. 4. En la figura se esquematiza el desempeño del clasificador con el conjunto de datos Phoneme.

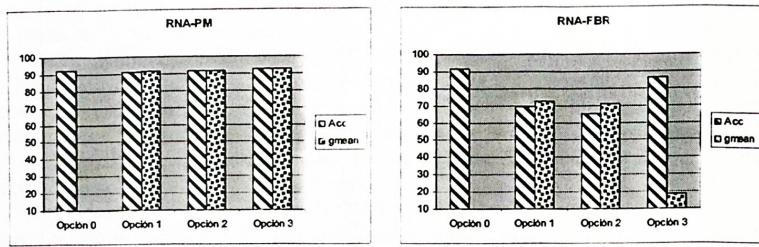


Fig. 5. En la figura se ilustra rendimiento del clasificador sobre la base de datos B2C1s.

#### 4 Conclusión

Los resultados anteriores nos dan una idea de las diferentes situaciones que pueden ocurrir cuando se trabaja con RNA entrenadas con el algoritmo back-propagation con procesamiento por grupos y bases de datos desbalanceadas de dos clases.

Se evidencia que la RNA-FBR es la más sensible a la presencia de desbalance en la ME. Además, se observan beneficios en términos de convergencia al aplicar funciones de costo para evitar que las clases menos representadas sean ignoradas en el proceso de entrenamiento (en particular en la RNA-PM). En este mismo sentido, se pueden ver incrementos considerables en los valores de  $g$ -mean, es decir, ayudan a incrementar el porcentaje de aciertos de las clases minoritarias.

Es indudable que se requiere ampliar y profundizar en los estudios mencionados, no tan sólo por la importancia del tema sino también por su vínculo con otras tareas del aprendizaje automático y del reconocimiento de patrones.

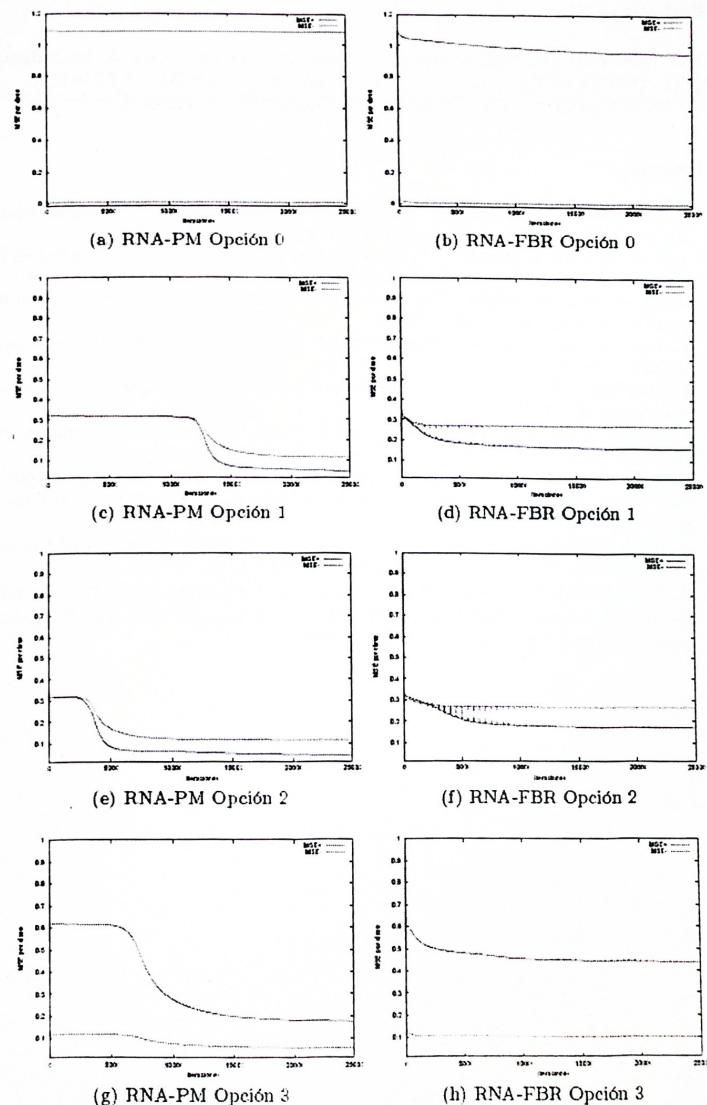


Fig. 6. En la figura se ilustra el MSE por clase de la base de datos B2Cls.

### Agradecimientos

Este trabajo ha sido parcialmente financiado por los proyectos de investigación DPI2006-15542-C04-03 del CICYT (España), SEP-2003-C02-44225 del CONACYT (Méjico), y GV/2007/105 de la Generalitat Valenciana.

### Referencias

1. R. Barandela, R.M. Valdovinos, J.S. Sánchez, and F.J. Ferri. The imbalanced training sample problem: Under or over sampling? In SSPR/SPR, page 806, 2004.
2. S. Kotsiantis and P. Pintelas. Mixture of expert agents for handling imbalanced data sets. *Annals of Mathematics and Computing & Tele-Informatics*, 1(1):46–55, 2003.
3. N. Japkowicz and S. Stephen. The class imbalance problem: a systematic study. *Intelligent Data Analysis*, 6:429–449, 2002.
4. T. Fawcett and F. Provost. Adaptive fraud detection. *Data Min. Knowl. Discov.*, 1(3):291–316, 1997.
5. Y. Murphrey, H. Guo, and L. Feldkamp. Neural learning from unbalanced data. *Applied Intelligence*, 21:117–128, 2004.
6. P.K. Chan, W. Fan, A.L. Prodromidis, and S.J. Stolfo. Distributed data mining in credit card fraud detection. *IEEE Intelligent Systems*, 14(6):67–74, 1999.
7. Z.-H. Zhou and X.-Y. Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18:63–77, 2006.
8. L. Bruzzone and S.B. Serpico. Classification of imbalanced remote-sensing data by neural networks. *Pattern Recognition Letters*, 18:1323–1328, 1997.
9. R. Anand, K.G. Mehrotra, C.K. Mohan, and S. Ranka. An improved algorithm for neural network classification of imbalanced training sets. *IEEE Transactions on Neural Networks*, 4:962–969, 1993.