Psicothema
2000. Vol. 12, nº 3, pp. 503-510

ISSN 0214 - 9915 CODEN PSOTEG


# Análisis de datos faltantes mediante redes neuronales artificiales

José Blas Navarro Pastor y Josep María Losilla Vidal
Universidad Autónoma de Barcelona

En los últimos años se ha consolidado el uso de redes neuronales artificiales como complemento a los métodos estadísticos. Sin embargo, no se ha profundizado en el estudio de cómo las redes neuronales artificiales se ven afectadas por la presencia de datos faltantes, ni en el establecimiento de las mejores estrategias para abordarlos durante la fase de análisis estadístico. En nuestro trabajo investigamos la eficacia de diversas técnicas para afrontar los datos faltantes en análisis descriptivos univariantes y en la generación de modelos de clasificación, entre las que se incluyen redes neuronales del tipo perceptrón multicapa y de función base radial. Nuestros resultados sugieren que, en general, para los tipos de variables estudiados las redes neuronales artificiales son más eficaces en la disminución del error de imputación que otras técnicas de análisis ampliamente utilizadas cuando existe un nivel de correlación no nulo con otras variables registradas.

*Analysis of missing data with artificial neural networks: A simulation study.* In the last years it has been consolidated the use of artificial neural nets as a complement to statistical methods. However, it has not been deeply studied neither how the presence of missing data affects artificial neuronal nets nor the establishment of the best strategies to treat missing data in the stage of statistical analysis. In our work we investigate the effectiveness of different techniques to face missing data in univariant descriptive analysis and in the generation of classification models, including multilayer perceptron and radial basis function neural nets. Our results suggest that, in general, artificial neural nets are more effective in decreasing the imputation error than other broadly used analysis techniques.

La calidad de los datos es un concepto amplio que engloba diferentes aspectos y cuya evaluación va adquiriendo mayor relevancia en la investigación psicológica en los últimos años (Doménech, Losilla y Portell, 1998; Freedland y Carney, 1992). Redman (1992) enumera cuatro dimensiones que hacen referencia al valor que pueden tomar los datos: precisión, actualidad, consistencia y plenitud. En lo referente a la plenitud de los datos, o lo que es lo mismo, al grado en que los datos a analizar contienen datos faltantes, siguiendo las propuestas de Azorín y Sánchez-Crespo (1986) y de Groves (1989), diferenciamos tres tipos de ausencia de información: los errores de cobertura, la falta de respuesta total y la falta de respuesta parcial. Los errores de cobertura y la falta de respuesta total deben ser abordados durante las fases de selección de la muestra y obtención de los datos, respectivamente, mientras que la falta de respuesta parcial se puede tratar durante la fase de análisis.

Desde la estadística se han propuesto diferentes estrategias de análisis de datos faltantes, como la imputación directa de la media, moda, o mediana, la imputación por regresión (Buck, 1960), la imputación múltiple (Schafer, 1997) o el algoritmo expectativa-maximización (algoritmo EM) (Dempster, Laird y Rubin, 1977). Consideradas globalmente, se trata de técnicas que palian en parte el problema de la falta de información, si bien no se pueden considerar soluciones óptimas por dos motivos principales: (1) pueden provocar un importante sesgo en las posteriores estimaciones realizadas sobre los datos completos, especialmente de variancias y covariancias, y (2) requieren el cumplimiento de determinados supuestos sobre la distribución de los datos (Little y Rubin, 1987).

Otra estrategia de análisis de datos que contienen valores faltantes, ampliamente utilizada en la práctica e implementada como opción por defecto en los paquetes estadísticos de uso más habitual, consiste en eliminar del análisis los registros que presentan algún valor faltante (método *listwise*). Como señalan Graham, Hofer y Piccinin (1994), Huberty y Julian (1995), Orme y Reis (1991), Smith (1991) y Whitehead (1994), entre otros, este método implica una sustancial mengua del tamaño muestral, que conlleva a su vez una reducción en la precisión de las estimaciones de los parámetros poblacionales y, además, una disminución de la potencia de las pruebas estadísticas de significación, por lo que en general es desaconsejable.

En los últimos años también se ha abordado el problema de los datos faltantes mediante redes neuronales artificiales (RNA). Las RNA se definen como un sistema de procesamiento de información, formado por un conjunto de unidades simples o procesadores organizadas en paralelo, que operan sólo con la información disponible localmente que reciben a través de las conexiones con otras unidades por las que fluye información de tipo numérico (Rumelhart y McClelland, 1986). Una tipología de RNA que se emplea habitualmente en la generación de modelos de clasificación y predicción son las denominadas RNA supervisadas, entre las cuales destacan, tanto por el número de trabajos que las utili-

Correspondencia: José Blas Navarro Pastor
Facultad de Psicología
Universidad Autónoma de Barcelona
08193 Barcelona (Spain)
E-mail: joseblas.navarro@uab.es

zan como por su amplia aplicabilidad, las redes perceptrón multicapa (MLP), consideradas aproximadores universales de funciones (Hornik, Stinchcombe y White, 1989, 1990). No nos detendremos aquí a exponer los detalles de la arquitectura de este tipo de RNA, ya que se trata de una línea de investigación consolidada; buena prueba de ello son las excelentes revisiones de Hecht-Nielsen (1989), Sarle (1994) y White (1992) y, en castellano, de Hilera y Martínez (1995) y Martín y Sanz (1997).

Actualmente las RNA son ampliamente utilizadas en la generación de modelos de clasificación, constatándose en diversos estudios su superioridad frente a otras técnicas de clasificación. Así, por ejemplo, Bonelli y Parodi (1991), Huang y Lippman (1987), Sethi y Otten (1990) compararon modelos tradicionales y RNA, obteniendo resultados ligeramente favorables a estas últimas; Garson (1991) confirmó la superioridad de las redes frente a los modelos de regresión, *path analysis* y análisis discriminante; Schrodt (1993) determinó que las RNA son más eficaces que los algoritmos de inducción de reglas y el análisis discriminante; Ripley (1993) dedujo que las redes MLP son equivalentes a complejas técnicas estadísticas como la *projection pursuit regression;* Furlanello, Giuliani y Trentin (1995) corroboraron la mejor capacidad de discriminación de las redes RBF frente a la de modelos lineales multivariables; finalmente, Pitarque, Roy y Ruíz (1998) concluyen recientemente que las RNA son superiores a los modelos estadísticos clásicos en la generación de modelos de clasificación.

También en esta línea se han publicado trabajos específicos que analizan el efecto de los valores faltantes en el procesamiento de la información que llevan a cabo las RNA, y se han sugerido diversas estrategias para imputar este tipo de valores (Pitarque y Ruiz, 1996; Sharpe y Solly, 1995; Tresp, Ahmad y Neuneier, 1997; Vamplew y Adams, 1992), algunas de las cuales han sido aplicadas con éxito en la clasificación de sujetos psicopatológicos (Pitarque, Ruiz, Fuentes, Martínez y García-Merita, 1997). Sin embargo, en la literatura revisada no hemos hallado ningún trabajo que evalúe los diferentes métodos de imputación en función del tipo de variable y del nivel de correlación de los datos. Sólo Prechelt (1994) hace referencia a la naturaleza de la variable en el momento de codificar los valores faltantes en las variables predictoras de una RNA, si bien, al tratarse de una primera comunicación a una lista de correo, plantea más preguntas que respuestas ofrece.

El objetivo general de este trabajo es comparar diferentes técnicas de afrontamiento de los datos faltantes en análisis descriptivos univariados y en la generación de modelos de clasificación, aportando sugerencias específicas en función de diversas características de los datos que se analizan, e incluyendo como estrategia de análisis la imputación de valor mediante RNA del tipo MLP y también del tipo función base radial (RBF), similares a las MLP por lo que respecta a su estructura y dirección del flujo de la información entre unidades pero que, como señala Orr (1996), difieren fundamentalmente de las MLP en la forma de establecer el valor de los parámetros, lo cual reduce el tiempo necesario para el aprendizaje aunque a costa de reducir también su aplicabilidad en problemas en los que se maneja una gran cantidad de valores de entrada.

Los objetivos específicos que se derivan del objetivo general son:

*Objetivo 1*: Comparar diferentes métodos de imputación de valor a los datos faltantes de una variable, teniendo en cuenta la naturaleza de la variable y su nivel de asociación con otras variables registradas.

*Objetivo 2:* Comparar diferentes métodos de tratamiento de los valores faltantes de las variables predictoras en la generación de modelos de clasificación de una variable criterio binaria con distribución equiprobable, teniendo en cuenta la técnica de clasificación empleada y el nivel de asociación entre la variable criterio y sus predictoras.

*Objetivo 3:* Investigar cómo incide el porcentaje de valores faltantes en las conclusiones del objetivo 2.

## Método

### *Procedimiento*

Para conseguir los objetivos planteados realizamos un experimento de simulación estocástica, en el que generamos inicialmente 300 matrices de datos completos (MC_1 a MC_300) con 500 registros y 7 variables cada una:

– Binaria con distribución equiprobable (BU) y no equiprobable (BA).

– Ordinal con tres categorías con distribución equiprobable (OU) y no equiprobable (OA).

– Cuantitativa con distribución normal (CN) y con distribución asimétrica (CA).

– Binaria con distribución equiprobable (DEP)

En la Tabla 1 se presentan los parámetros empleados en el mecanismo generador de datos para cada variable. En los posteriores modelos de clasificación, las seis primeras variables actuarán como predictoras y la séptima como criterio.

El nivel de correlación tanto entre las variables predictoras como entre éstas y la criterio se manipuló para crear tres grupos de matrices, cada uno con un total de 100 matrices (ver Tabla 2).

A partir de cada matriz completa generamos 2 nuevas matrices incompletas (MM1 y MM2). Para crear las matrices MM1 eliminamos, mediante un procedimiento aleatorio, 15 valores en cada una de las 6 primeras variables, de manera que el porcentaje de valores faltantes fuera del 3%. En las matrices MM2 el porcentaje de valores faltantes se incrementó hasta el 12%.

Para conseguir el objetivo 1, los valores faltantes de cada variable en las matrices MM1 fueron imputados mediante métodos de imputación directa, considerando como tal los que calculan el valor a imputar empleando únicamente la información de la variable incompleta, y métodos de imputación por regresión/red, incluyendo en este grupo las técnicas que calculan el valor a imputar a partir de los datos de la variable incompleta y del resto de variables registradas. Los métodos de imputación directa analizados son:

*Tabla 1*
Parámetros empleados para generar cada variable

| Predictoras | | | | | | Criterio |
|---|---|---|---|---|---|---|
| **BU** | **BA** | **OU** | **OA** | **CN** | **CA*** | **DEP** |
| $\pi$=0.5 | $\pi$=0.8 | $\pi_1$=1/3<br>$\pi_2$=1/3<br>$\pi_3$=1/3 | $\pi_1$=0.50<br>$\pi_2$=0.35<br>$\pi_3$=0.15 | $\mu$=1<br>$\sigma$=0 | $\mu$=1.6<br>$\sigma$=2.1 | $\pi$=0.5 |

* Coeficiente de asimetría=4.8, Error estándar=0.11

BU: Binaria equiprobable; BA: Binaria no equiprobable; OU: Ordinal equiprobable; OA: Ordinal no equiprobable; CN: Cuantitativa normal; CA: Cuantitativa asimétrica; DEP: Dependiente

– Imputación de la media aritmética (MEDIA) a las variables cuantitativas y de la moda (MODA) a las categóricas.

– Imputación de la mediana (MDNA).

– Imputación de un valor aleatorio procedente de una distribución de probabilidad uniforme, dentro del rango de valores observados (VADU).

– Imputación de un valor aleatorio procedente de una distribución de probabilidad estimada para cada tipo de variable, dentro del rango de valores observados (VADE).

Los métodos de imputación por regresión/red analizados son:

– Imputación del valor predicho por un modelo de regresión lineal múltiple (variables cuantitativas) o de regresión logística (variables categóricas) (REG).

– Imputación mediante el algoritmo EM (variables cuantitativas) (EM).

– Imputación del valor predicho por una RNA perceptrón multicapa (MLP).

– Imputación del valor predicho por una RNA de función base radial (RBF).

En cada caso medimos el error de imputación promedio cometido. Para calcular el error de imputación, en las variables binarias y ordinales se calculó la tasa nominal de error (TNE), considerando el error igual a 0 si el valor imputado coincidía con el valor real e igual a 1 en caso contrario, y promediando dichos valores entre todos los registros incompletos. En las variables cuantitativas se calculó el error medio cuadrático (EMC), promediando la suma de las diferencias al cuadrado entre el valor imputado y el valor real.

Puesto que el resultado de los métodos de imputación directa no está afectado por la magnitud de la asociación entre variables, estos métodos sólo fueron evaluados en las matrices MM1_1 a MM1_100. Respecto a los métodos de imputación por regresión/red, hay que tener en cuenta que los valores faltantes en las variables que actúan como predictoras en el modelo de imputación fueron temporalmente reemplazados con el resultado del mejor método de imputación directa.

Para conseguir el objetivo 2 aplicamos a las matrices MM1 los siguientes métodos de tratamiento de los valores faltantes:

– Eliminación de aquellos registros que presenten algún valor faltante (*Listwise*).

– Imputación de todos los valores faltantes mediante el mejor método de imputación directa, seleccionado en el objetivo 1, para cada tipo de variable estudiado (MID).

– Codificación de todos los valores faltantes al valor 99 (CODI). Diferentes análisis realizados previamente por los autores pusieron de manifiesto que el valor empleado para codificar los datos ausentes (-99, 99, 999) no es relevante en el resultado de la clasificación, siempre y cuando sea un valor fuera del rango de valores observados de la variable.

– Codificación de todos los valores faltantes al valor 99 y generación de una variable indicadora por cada variable predictora (COVI). La variable indicadora tendrá el valor 1 si el dato original es desconocido y el valor 0 en caso contrario.

En las matrices con datos completos resultantes estimamos diferentes modelos de clasificación de la variable criterio, empleando para ello las siguientes técnicas de clasificación:

– Regresión logística (RL).

– RNA perceptrón multicapa (MLP).

– RNA de función base radial (RBF).

El error de clasificación en cada modelo se midió como el porcentaje de clasificaciones incorrectas realizadas en una submuestra aleatoria de 50 registros que actúan como datos de test. Con el objetivo de disponer de un error de clasificación de referencia también se estimaron modelos de clasificación en las matrices de datos completos originales (MC).

Para conseguir el objetivo 3 se estimaron nuevos modelos de clasificación aplicando los métodos *listwise*, CODI, COVI y MID a las matrices MM2. Únicamente trabajamos con las matrices que presentan correlación no nula entre las variables predictoras y la criterio (matrices MM2_101 a MM2_300), ya que, si un incremento del porcentaje de valores faltantes repercute sobre el resultado de la clasificación, sólo será apreciable cuando los datos estén relacionados.

*Material*

La generación de las matrices de datos completos y con valores faltantes se realizó mediante el programa Matlab 4.1 (The MathWorks Inc., 1994). Para llevar a cabo los diferentes métodos de imputación directa y por regresión/red de los valores faltantes, así como la estimación de los modelos de clasificación empleamos el módulo de análisis de datos perdidos de la aplicación SPSS en su versión 7.5.2. para Windows (SPSS Inc., 1996), el módulo de regresión logística del BMDP 93' (Dixon, 1993) y la aplicación Neural Connection 2.0 (SPSS Inc., 1997).

Para decidir la topología de RNA a emplear en los diferentes apartados de la simulación se realizaron una serie de pruebas previas con un subconjunto de las matrices disponibles. Partiendo siempre de la topología sugerida por la aplicación Neural Connection, se modificaron tanto el número de unidades ocultas (aumentándolo y disminuyéndolo), como el número de capas de unidades ocultas (una o dos capas). Los resultados demostraron que las diferencias entre las diferentes topologías probadas eran mínimas y no sistemáticas. A raíz de ello, siempre se trabajó con los parámetros sugeridos por la aplicación Neural Connection. Las redes MLP tenían una capa con 4 unidades ocultas en las redes más simples y 7 en las más complejas (éstas últimas empleadas sólo en el problema de clasificación mediante la técnica de codificación con inclusión de variables indicadoras), función de activación identidad para las unidades de entrada y de salida y

*Tabla 2*
Nivel de correlación entre variables predictoras y criterio

| | MC_1 a MC_100 | | MC_101 a MC_200 | | MC_201 a MC_300 | |
|---|---|---|---|---|---|---|
| | **V.P.** | **V.C.** | **V.P.** | **V.C.** | **V.P.** | **V.C.** |
| V.P. | Nulo<br>$r < 0.02$ | Nulo<br>$r < 0.02$ | Bajo<br>$0.02 \leq r < 0.10$ | Bajo-medio<br>$0.02 \leq r < 0.40$ | Medio-alto<br>$0.10 \leq r < 0.50$ | Alto<br>$0.40 \leq r < 0.50$ |

V.P.: Variables predictoras; V.C.: Variable criterio ; r: coeficiente de correlación lineal

logística para las unidades ocultas, y regla de aprendizaje de segundo orden denominada gradiente conjugado (Battiti, 1992). Los pesos iniciales de las conexiones se obtuvieron de forma aleatoria a partir de una distribución uniforme con rango -0.1 a 0.1. Para acelerar la convergencia, la aplicación Neural Connection realiza el entrenamiento de la RNA en 4 fases diferentes, iniciando el entrenamiento en las fases 2, 3 y 4 con la configuración de la red que proporciona el menor error de generalización en la fase anterior. Los coeficientes de aprendizaje y momento, junto al número de iteraciones (épocas) de cada fase de entrenamiento se hallan en la Tabla 3. El número de iteraciones en la última fase es superior para conseguir el ajuste óptimo a partir de los valores de los parámetros establecidos en las fases anteriores.

Las redes RBF se configuraron con 5 centros iniciales posicionados aleatoriamente, con incrementos de 5 en 5 hasta un máximo de 50 centros. La medida de distancia de error empleada fue la distancia euclídea y la función radial la función gausiana, con $\sigma^2$=0.1

## Resultados

El análisis de resultados que presentamos a continuación se centra en la evaluación del error de imputación (objetivo 1) y de clasificación (objetivos 2 y 3) cometido en cada condición planteada en el experimento de simulación. Debido a que en un experimento de simulación el número de muestras simuladas es fijado por los investigadores en un número elevado para obtener estimaciones precisas de los efectos de interés, para comparar los resultados obtenidos en las diversas condiciones simuladas considera-

*Tabla 3*
Características de la regla de aprendizaje de las redes MLP

| | Fase 1 | Fase 2 | Fase 3 | Fase 4 |
|---|---|---|---|---|
| Nº Iteraciones | 100 | 100 | 100 | 10000 |
| Coef. aprendizaje | 0.9 | 0.7 | 0.5 | 0.4 |
| Momento | 0.1 | 0.4 | 0.5 | 0.6 |

*Tabla 4*
Error de imputación (intervalo de confianza del 95%) cometido según técnica de imputación, nivel de correlación lineal y tipo de variable, en las matrices MM1

| | | BU* | BA* | OU* | OA* | CN** | CA** |
|---|---|---|---|---|---|---|---|
| MM1_1 a MM1_100 | MODA | 0.477 (0.453-0.501) | 0.204 (0.185-0.223) | 0.641 (0.616-0.666) | 0.497 (0.471-0.523) | N.A. | N.A |
| | MEDIA | N.A. | N.A. | N.A. | N.A. | 1.052 (0.967-1.137) | 5.329 (3.191-7.467) |
| | MDNA | N.A. | N.A. | 0.673 (0.647-0.699) | 0.566 (0.535-0.597) | 1.053 (0.968-1.138) | 5.720 (3.457-7.983) |
| | VADU | 0.475 (0.452-0.499) | 0.537 (0.513-0.561) | 0.661 (0.635-0.688) | 0.661 (0.637-0.686) | 4.302 (3.995-4.608) | 189.186 (142.3-236.0) |
| | VADE | 0.524 (0.500-0.548) | 0.291 (0.270-0.312) | 0.661 (0.635-0.688) | 0.604 (0.580-0.628) | 1.926 (1.789-2.064) | 11.939 (6.414-17.46) |
| | REG | 0.492 (0.466-0.518) | 0.205 (0.186-0.223) | 0.665 (0.641-0.689) | 0.509 (0.484-0.533) | 1.063 (0.977-1.149) | 5.352 (3.208-7.496) |
| | EM | N.A. | N.A. | N.A. | N.A. | 1.057 (0.972-1.143) | 5.350 (3.203-7.496) |
| | MLP | 0.511 (0.482-0.539) | 0.217 (0.198-0.235) | 0.644 (0.619-0.668) | 0.531 (0.505-0.556) | 1.088 (1.000-1.175) | 5.432 (3.297-7.566) |
| | RBF | 0.515 (0.488-0.541) | 0.210 (0.192-0.228) | 0.655 (0.632-0.679) | 0.519 (0.493-0.544) | 1.086 (0.996-1.177) | 5.507 (3.366-7.648) |
| MM1_101 a MM1_200 | REG | 0.448 (0.424-0.473) | 0.185 (0.164-0.206) | 0.633 (0.608-0.659) | 0.502 (0.478-0.526) | 0.978 (0.909-1.047) | 5.060 (4.380-5.739) |
| | EM | N.A. | N.A. | N.A. | N.A. | 0.967 (0.897-1.037) | 5.070 (4.396-5.743) |
| | MLP | 0.442 (0.415-0.469) | 0.212 (0.191-0.233) | 0.603 (0.575-0.630) | 0.527 (0.502-0.551) | 0.942 (0.872-1.012) | 5.153 (4.459-5.846) |
| | RBF | 0.424 (0.397-0.451) | 0.187 (0.166-0.208) | 0.600 (0.575-0.625) | 0.519 (0.493-0.545) | 0.923 (0.853-0.992) | 5.070 (4.369-5.771) |
| MM1_201 a MM1_300 | REGR | 0.313 (0.288-0.337) | 0.185 (0.166-0.203) | 0.477 (0.447-0.506) | 0.414 (0.391-0.437) | 0.680 (0.629-0.731) | 3.815 (2.632-4.997) |
| | EM | N.A. | N.A. | N.A. | N.A. | 0.637 (0.586-0.687) | 3.858 (2.658-5.058) |
| | MLP | 0.273 (0.252-0.295) | 0.194 (0.174-0.214) | 0.454 (0.429-0.479) | 0.429 (0.407-0.452) | 0.611 (0.556-0.666) | 4.112 (2.859-5.364) |
| | RBF | 0.279 (0.255-0.303) | 0.188 (0.169-0.207) | 0.438 (0.414-0.462) | 0.419 (0.393-0.445) | 0.595 (0.542-0.649) | 3.836 (2.696-4.976) |

* Tasa nominal de error ; ** Error medio cuadrático ; N.A.: No aplicado

BU: Binaria equiprobable; BA: Binaria no equiprobable; OU: Ordinal equiprobable; OA: Ordinal no equiprobable; CN: Cuantitativa normal; CA: Cuantitativa asimétrica; MDNA: Mediana; VADU: Valor aleatorio distribución uniforme; VADE: Valor aleatorio distribución estimada; REG: Regresión; EM: Algoritmo EM; MLP: Red perceptrón multicapa; RBF: Red de función base radial

mos más adecuada la interpretación de los intervalos de confianza de las medidas de error que el grado de significación estadística de la diferencia entre ellas, índice éste que no aporta información sobre la relevancia práctica de dichas diferencias. Por este motivo en todas las tablas de resultados se incluyen los límites de los intervalos con un nivel de confianza del 95%.

*Objetivo 1*: En la Tabla 4 se presenta el error cometido con cada técnica de imputación evaluada según el tipo de variable y el nivel de correlación lineal entre variables. Cuando la correlación es nula (matrices MM1_1 a MM1_100) los métodos de imputación directa y por regresión/red tienen una eficacia similar.

Si el nivel de correlación se incrementa hasta valores medios-altos (matrices MM1_201 a MM1_300), los cuatro métodos de imputación por regresión/red son claramente superiores a los de imputación directa. Concretamente, en las variables categóricas con distribución equiprobable (BU, OU) la imputación mediante red neuronal MLP o RBF ofrece los mejores resultados, mientras que en las de distribución no equiprobable (BA, OA) no se aprecian diferencias relevantes entre los tres métodos aplicados. En cuanto a las variables cuantitativas, en la variable CN el menor error se obtiene con red RBF, y en la variable CA los cuatro procedimientos ofrecen un error similar (ligeramente superior con red MLP).

Con niveles bajos de correlación (matrices MM1_101 a MM1_200) en general se obtienen resultados intermedios a los presentados anteriormente. La excepción se halla en las variables categóricas no equiprobables (BA, OA), en las que el error no se reduce respecto a las matrices con ausencia de correlación, y en especial en BA, en la que ni siquiera se incrementa respecto a las matrices con correlación media-alta.

*Objetivo 2:* En la Tabla 5 se presenta el error de clasificación cometido con cada técnica estudiada según la estrategia de afrontamiento de los valores faltantes y el nivel de correlación lineal entre la variable criterio y sus predictoras, utilizando el conjunto de matrices MM1 (3% de valores faltantes). Para facilitar la comparación también se incluyen los resultados obtenidos con las matrices de datos completos (MC). En las matrices con correlación nula (MM1_1 a MM1_100) el error de clasificación es muy similar en todas las condiciones planteadas, si bien cabe mencionar que la clasificación mediante red neuronal MLP o RBF, habiendo eliminado los registros con datos faltantes (*listwise*), proporciona el menor error. Los porcentajes de error son muy elevados si se tiene en cuenta que por azar se clasificarían correctamente un 50% de registros.

Al incrementar la correlación hasta niveles altos (MM1_201 a MM1_300) se manifiestan importantes diferencias. Un primer análisis refleja que la clasificación mediante red MLP reduce alrededor de un 9% el porcentaje de clasificaciones incorrectas respecto a la regresión logística, mientras que en la clasificación con red RBF la reducción se sitúa en torno al 8%. Respecto a los mé-

*Tabla 5*
Error de clasificación (intervalo de confianza del 95%) cometido según técnica de clasificación, nivel de correlación lineal y método de tratamiento de los valores faltantes, en las matrices MM1

| | | RL | MLP | RBF |
|---|---|---|---|---|
| MM1_1 a MM1_100 | LISTWISE | 0.443 (0.439-0.447) | 0.404 (0.392-0.415) | 0.408 (0.401-0.415) |
| | MID | 0.447 (0.443-0.451) | 0.436 (0.426-0.445) | 0.423 (0.416-0.429) |
| | CODI | 0.440 (0.437-0.443) | 0.426 (0.417-0.434) | 0.426 (0.419-0.432) |
| | COVI | 0.436 (0.433-0.440) | 0.419 (0.409-0.428) | 0.431 (0.424-0.437) |
| | MC | 0.446 (0.443-0.450) | 0.425 (0.415-0.435) | 0.423 (0.417-0.429) |
| MM1_101 a MM1_200 | LISTWISE | 0.225 (0.222-0.229) | 0.160 (0.154-0.166) | 0.165 (0.160-0.171) |
| | MID | 0.234 (0.231-0.237) | 0.169 (0.164-0.175) | 0.179 (0.173-0.185) |
| | CODI | 0.282 (0.278-0.287) | 0.192 (0.185-0.200) | 0.222 (0.200-0.214) |
| | COVI | 0.229 (0.226-0.232) | 0.169 (0.163-0.176) | 0.207 (0.164-0.175) |
| | MC | 0.217 (0.214-0.220) | 0.162 (0.158-0.167) | 0.169 (0.164-0.175) |
| MM1_201 a MM1_300 | LISTWISE | 0.212 (0.208-0.216) | 0.118 (0.111-0.126) | 0.130 (0.126-0.135) |
| | MID | 0.215 (0.212-0.219) | 0.131 (0.126-0.137) | 0.137 (0.132-0.142) |
| | CODI | 0.232 (0.229-0.236) | 0.144 (0.138-0.149) | 0.150 (0.145-0.155) |
| | COVI | 0.210 (0.207-0.214) | 0.136 (0.130-0.143) | 0.149 (0.144-0.154) |
| | MC | 0.213 (0.209-0.216) | 0.129 (0.123-0.134) | 0.135 (0.130-0.139) |

RL: Regresión logística; MLP: Red perceptrón multicapa; RBF: Red de función base radial; MID: Mejor imputación directa; CODI: Codificación; COVI: Codificación con variables indicadoras; MC: Matriz completa

todos de tratamiento de los valores faltantes, la eliminación de los registros incompletos (*listwise*) seguido de la imputación directa (MID) ofrecen el menor error.

En las matrices con correlación baja-media (MM1_101 a MM1_200) de nuevo se constata la superioridad de las RNA, y en especial del tipo MLP, respecto a la regresión logística. Centrándonos en los datos obtenidos mediante la clasificación con red MLP, el error de clasificación es claramente superior con el procedimiento de codificación (CODI), en tanto que la eliminación de registros *(listwise)*, seguido de la imputación directa (MID) y la codificación con inclusión de variables indicadoras (COVI) ofrecen el menor error.

*Objetivo 3:* En la Tabla 6 se presenta el error de clasificación cometido con cada técnica estudiada utilizando el conjunto de matrices MM2_101 a MM2_300 (12% de valores faltantes y correlación no nula). Para facilitar la comparación se incluye también el resultado de la clasificación en las matrices completas (MC).

La técnica de eliminación de registros ofrece de nuevo el menor error de clasificación, que prácticamente coincide con el obtenido en las matrices de datos con un 3% de valores faltantes. Dejando de lado el método *listwise*, todos los porcentajes de error obtenidos en las matrices MM2 son superiores a los correspondientes calculados en las matrices MM1. Un estudio más detallado pone de manifiesto que con la imputación directa (MID) y la codificación con inclusión de variables indicadoras (COVI) se consiguen los mejores resultados.

## Discusión y conclusiones

La conclusión general que se desprende de nuestros resultados es que la estrategia óptima para el análisis de datos faltantes depende de la naturaleza de la variable estudiada y de su nivel de correlación con otras variables registradas.

En lo referente a análisis univariantes, la Tabla 7 presenta el método de imputación recomendable en base a los resultados de nuestro estudio. Si en una celda de la tabla hay más de un método todos ellos se consideran similares, a pesar de lo cual el orden en que aparecen refleja nuestra preferencia. Además del error de imputación, en dicha elección hemos considerado la sencillez de realización y que las RNA no establecen ningún tipo de supuesto sobre la distribución de los datos, mientras que el modelo de regresión y el algoritmo EM requieren la comprobación empírica de que la distribución poblacional de las variables se ajusta a un determinado modelo probabilístico, aspecto que, como afirman Graham, Hofer y Mackinnon (1996) difícilmente puede ser corroborado a partir de los datos disponibles en una muestra.

Respecto a la generación de un modelo de clasificación, la estrategia de afrontamiento de los valores faltantes que ofrece el menor error es la eliminación de los registros que tienen algún valor faltante (*listwise*). Sin embargo, como hemos comentado en la introducción, este método disminuye el tamaño muestral disponible, la precisión de las estimaciones y la potencia de las pruebas estadísticas de significación, por lo que, excepto cuando el número de registros con valores faltantes es muy bajo, es desaconsejable. A primera vista, nuestros resultados son diametralmente opuestos a los obtenidos por Pitarque y Ruiz (1996), quienes concluyen que con la técnica *listwise* se obtiene el peor error de clasificación. Sin embargo, las condiciones de la simulación realizada permiten explicar las discrepancias. Pitarque y Ruiz (1996) trabajan con matrices de datos con 15 variables predictoras y 100 registros, de los cuales la mitad tienen dos valores ausentes, de manera que al eliminar los registros incompletos disponen de 50 casos para estimar 103 parámetros (implicados en la red MLP 15-6-1 que emplean). Por contra, nuestras matrices de datos contienen 500 registros y 6 variables predictoras, con un 3% de valores faltantes en cada va-

*Tabla 6*
Error de clasificación (intervalo de confianza del 95%) cometido según técnica de clasificación, nivel de correlación lineal y método de tratamiento de los valores faltantes, en las matrices MM2

| | | RL | MLP | RBF |
|---|---|---|---|---|
| MM2_101 a MM2_200 | LISTWISE | 0.228 (0.222-0.234) | 0.165 (0.158-0.172) | 0.167 (0.160-0.174) |
| | MID | 0.250 (0.246-0.254) | 0.181 (0.175-0.187) | 0.189 (0.183-0.196) |
| | CODI | 0.311 (0.308-0.314) | 0.198 (0.193-0.202) | 0.227 (0.223-0.231) |
| | COVI | 0.256 (0.253-0.259) | 0.180 (0.176-0.184) | 0.215 (0.210-0.220) |
| | MC | 0.217 (0.214-0.220) | 0.162 (0.158-0.167) | 0.169 (0.164-0.175) |
| MM2_201 a MM2_300 | LISTWISE | 0.207 (0.202-0.213) | 0.111 (0.102-0.119) | 0.134 (0.127-0.141) |
| | MID | 0.233 (0.229-0.236) | 0.147 (0.143-0.152) | 0.150 (0.145-0.155) |
| | CODI | 0.263 (0.259-0.267) | 0.153 (0.147-0.159) | 0.159 (0.154-0.164) |
| | COVI | 0.246 (0.243-0.250) | 0.148 (0.142-0.153) | 0.150 (0.145-0.155) |
| | MC | 0.213 (0.209-0.216) | 0.129 (0.123-0.134) | 0.135 (0.130-0.139) |

RL: Regresión logística; MLP: Red perceptrón multicapa; RBF: Red de función base radial; MID: Mejor imputación directa; CODI: Codificación; COVI: Codificación con variables indicadoras; MC: Matriz completa

riable. En el peor de los casos (ningún registro con 2 valores faltantes), ello se traduce en la existencia de 90 registros con datos faltantes, lo que supone que quedan 410 casos para estimar 57 parámetros (implicados en la red MLP 12-4-1 que empleamos en la mayoría de modelos).

Dejando de lado el método *listwise*, los métodos recomendados son la imputación directa de los datos faltantes de cada variable (MID) y la codificación de los valores faltantes con la inclusión de variables indicadoras (COVI), estimando el posterior modelo de clasificación mediante RNA del tipo MLP. Estos resultados coinciden con los hallados por Vamplew y Adams (1992), quienes obtienen los mejores porcentajes de clasificación con la inclusión de variables indicadoras y mediante la imputación de los valores faltantes con una RNA. Asimismo, en un contexto psicopatológico, con el objetivo de clasificar sujetos como patológicos o no a partir de la presencia/ausencia de determinados síntomas, Taylor y Amir (1994) deducen que el empleo de variables indicadoras es una solución óptima.

Los resultados respecto a la influencia del porcentaje de valores faltantes sugieren que, en general, la capacidad discriminante de un modelo de clasificación se reduce a medida que aumenta dicho porcentaje. También se aprecia una tendencia a que cuanto más alto es el nivel de correlación con la variable criterio, más se ve afectado el error de clasificación por un incremento del porcentaje de datos faltantes. Por último, la pérdida de capacidad predictiva es mayor cuando se clasifica mediante regresión logística que cuando se hace con RNA y, dentro de éstas, las redes MLP se ven algo más afectadas que las redes RBF. Todo ello hace sospechar que las RNA, y especialmente las redes RBF, son más resistentes al incremento del porcentaje de valores faltantes que la regresión logística.

*Tabla 7*
Método de imputación recomendado según nivel de covariación y tipo de variable

| | Nivel de correlación | | |
|---|---|---|---|
| | **Nulo** | **Bajo** | **Medio-alto** |
| BU | MODA | RBF | RBF / MLP |
| BA | MODA | RBF / REG | RBF / REG |
| OU | MODA | RBF / MLP | RBF / MLP |
| OA | MODA | REG | RBF / REG |
| CN | MEDIA | RBF / MLP | RBF / MLP |
| CA | MEDIA | RBF / REG /EM | RBF / REG / EM |

BU: Binaria equiprobable; BA: Binaria no equiprobable; OU: Ordinal equiprobable; OA: Ordinal no equiprobable; CN: Cuantitativa normal; CA: Cuantitativa asimétrica; REG: Regresión; EM: Algoritmo EM ; MLP: Red perceptrón multicapa; RBF: Red de función base radial

Por último, la estimación de diferentes modelos de clasificación con las matrices de datos completas nos permite concluir que las técnicas presentadas son efectivas cuando hay un número reducido de valores faltantes, mientras que, en caso contrario, cualquier procedimiento se muestra incapaz de compensar totalmente la ausencia de información provocada por la pérdida de datos.

A nuestro juicio el presente trabajo aporta una solución sencilla al habitual problema de la falta de información, y en dicha sencillez radica su principal virtud. El tradicional método *listwise*, descalificado por diferentes autores, y a pesar de ello empleado masivamente en la investigación actual, sólo puede ser relegado al olvido si su heredero goza de la única virtud que, indiscutiblemente, se le puede atribuir: simplicidad. Las redes neuronales artificiales no requieren un profundo conocimiento matemático previo a su uso aplicado y, además, cada día son más los programas informáticos que las incorporan. Otros procedimientos habituales, particularmente la imputación del valor medio o la moda, también poseen la cualidad de ser aplicados con suma simplicidad, pero en determinadas circunstancias conllevan sesgos tan importantes que su uso es más perjudicial que beneficioso. En este sentido, creemos que un aspecto relevante de nuestro trabajo consiste en seleccionar la mejor estrategia para imputar los valores faltantes en función del tipo de variable a imputar y del nivel de correlación en los datos. Consideramos, como se observa genéricamente en la Tabla 7, los métodos estadísticos y los modelos de red neuronal como técnicas complementarias, cada una de las cuales resultará de mayor o menor utilidad según el problema específico que se aborde.

## Agradecimientos

Este trabajo ha sido posible gracias a la ayuda DGICYT PM98-0173 del Ministerio de Educación y Cultura.

## Referencias

Azorín, F. y Sánchez-Crespo, J.L. (1986). *Métodos y aplicaciones del muestreo.* Madrid: Alianza Editorial.

Battiti, R. (1992). First and second order methods for learning: between steepest descent and Newton's method. *Neural Computation, 4*(2), 141-166.

Bonelli, P. y Parodi, A. (1991). An efficient classifier system and its experimental comparisons with two representative learning methods on three medical domains. *Proceedings of the International Conference on Genetic Algorithms*, 288-295.

Buck, S.F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society, B22*, 302-306.

Dempster, A.P., Laird, N.M. y Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B39*, 1-38.

Dixon, W. (1993). *Biomedical computer programs (BMDP)* [Programa para ordenador]. SPSS Inc. (Productor). Chicago: SPSS Inc. (Distribuidor).

Doménech, J.M., Losilla, J.M. y Portell, M. (1998). La verificació aleatòria: una estratègia per millorar i avaluar la qualitat de l'entrada de dades. *Qüestió, 22*(3), 493-510.

Freedland, K.E. y Carney, R.M. (1992). Data management and accountability in behavioral and biomedical research. *American Psychologist, 47*(5), 640-645.

Furlanello, C., Giuliani, D. y Trentin, E. (1995). Connectionist speaker normalization with generalized resource allocating networks. En G. Tesauro, D. Touretsky y T.K. Leen (Eds.). *Advances in neural information processing systems (NIPS)* (pp. 867-874). Cambridge: MIT Press.

Garson, G.D. (1991). A comparison of neural network and expert systems algorithms with common multivariate procedures for analysis of social science data. *Social Science Computer Review, 9*, 399-434.

Graham, J.W., Hofer, S.M. y Piccinin, A.M. (1994). Analisys with missing data in drug prevention research. En L.M. Collins y L.A. Seitz (Eds.).

*Advances in data analysis for prevention intervention research* (pp. 13-63). Washington, DC: National Institute on Drug Abuse.

Graham, J.W., Hofer, S.M. y Mackinnon, D.P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: an application of maximum likelihood procedures. *Multivariate Behavioral Research, 31*(2), 197-218.

Groves, R.M. (1989). *Survey errors and survey costs*. New York: John Wiley & Sons.

Hecht-Nielsen, R. (1989). Theory on the back-propagation neural network. *Proceedings of the International Joint Conference on Neural Networks, I*, 593-606.

Hilera, J.R. y Martínez, V.J. (1995). *Redes neuronales artificiales: fundamentos, modelos y aplicaciones*. Madrid: RA-MA.

Hornik, K., Stinchcombe, M. y White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*, 359-366.

Hornik, K., Stinchcombe, M. y White, H. (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks, 3*, 551-560.

Huang, W.Y. y Lippmann, R.P. (1987). Comparisons between neural net and conventional classifiers. *Proceedings of the IEEE International conference on neural networks, I*, 485-494.

Huberty, C.J. y Julian, M.W. (1995). An ad hoc analysis strategy with missing data. *Journal of Experimental Education, 63*, 333-342.

Little, R.J.A. y Rubin, D.B. (1987). *Statistical analysis with missing data*. Nueva York: John Wiley & Sons.

Martín, B. y Sanz, A. (1997). *Redes neuronales y sistemas borrosos*. RA-MA: Madrid.

Orme, J.G. y Reis, J. (1991). Multiple regression with missing data. *Journal of Social Service Research, 15*, 61-91.

Orr, M.J.L. (1996). *Introduction to radial basis functions*. Edimburgh: University of Edinburgh, Center for cognitive science.

Pitarque, A. y Ruiz, J.C. (1996). Encoding missing data in back-propagation neural networks. *Psicológica, 17*, 83-91.

Pitarque, A., Ruiz, J.C., Fuentes, I., Martínez, M.J. y García-Merita, M. (1997). Diagnóstico clínico en psicología a través de redes neuronales. *Psicothema, 9*, 359-363.

Pitarque, A., Roy, J.F. y Ruiz, J.C. (1998). Redes neuronales vs modelos estadísticos: simulaciones sobre tareas de predicción y clasificación. *Psicológica, 19*, 387-400.

Prechelt, L. (1994). *Encoding missing values* [Archivo de datos informático]. Acceso e-Mail: ml-connectionists-request@TELNET-1.SRV.CS.CMU.EDU Nombre de la lista: Connectionists. Emisor: prechelt@ira.uka.de.

Redman, T.C. (1992). *Data quality: management and technology*. New York: Bantam Books.

Ripley, B.D. (1993). Statistical aspects of neural networks. En O.E. Barndorff-Nielsen, J.L. Jensen y W.S. Kendall (Eds.). *Networks and chaos: statistical and probabilistic aspects*. Londres: Chapman and Hall.

Rumelhart, D.E. y McClelland, J.L. (Eds.). (1992). *Introducción al procesamiento distribuido en paralelo* (García, J.A., Trad.). Madrid: Alianza Editorial. (Traducción del original Paralell distributed processing, 1986).

Sarle, W.S. (1994). *Neural networks and statistical models*. Proceedings of the nineteenth anual SAS users group international conference, Massachusetts.

Schafer, J.L. (1997). *Analysis of incomplete multivariate data*. Londres: Chapman and Hall.

Schrodt, P.A. (1991). Prediction of interstate conflict outcomes using a neural network. *Social Science Computer Review, 9*, 359-380.

Sethi, I.K. y Otten, M. (1990). Comparison between entropy net and decision tree classifiers. *Proceedings of the International Joint Conference on Neural Networks, I*, 63-68.

Sharpe, P.K. y Solly, R.J. (1995). Dealing with missing values in neural network-based diagnostic systems. *Neural Computing and Applications, 3*, 73-77.

Smith, T.W. (1991). *An analysis of missing income information on the General Social Surveys* (Informe metodológico Nº 71), Universidad de Chicago. Acceso HTTP: http://www.icpsr.umich.edu/ gss/report/m-report/meth71.htm.

SPSS Inc. (1996). *Statistical Package for Social Sciences 7.5.2* [Programa para ordenador]. SPSS Inc. (Productor). Chicago: SPSS Inc. (Distribuidor).

SPSS Inc. (1997). *Neural Connection 2.0* [Programa para ordenador]. SPSS Inc. (Productor). Chicago: SPSS Inc. (Distribuidor).

Taylor, M.A. y Amir, N. (1994). The problem of missing clinical data for research in psychopathology. *The Journal of Nervous and Mental Disease, 182*, 222-229.

The MathWorks (1994). *Matlab 4.1* [Programa para ordenador]. The MathWorks Inc. (Productor). Natick, Mas: The MathWorks Inc. (Distribuidor).

Tresp, V., Ahmad, S. y Neuneier, R. (1994). Training neural networks with deficient data. En J.D. Cowan, G. Tesauro y J. Alspector (Eds.). *Advances in neural information processing systems (NIPS)*. San Mateo: Morgan Kaufmann.

Vamplew, P. y Adams, A. (1992). Missing values in a backpropagation neural net. *Proceedings of the 3rd. Australian Conference on Neural Networks (ACNN), I*, 64-66.

White, H. (1992). *Artificial neural networks: approximation and learning theory*. Oxford: Blackwell.

Whitehead, J.C. (1994). Item nonresponse in contingent valuation: Should CV researchers impute values for missing independent variables? *Journal of Leisure Research, 26*, 296-303.

*Aceptado el 23 de diciembre de 1999*