Imputación de datos con redes neuronales

María E. Valesani
Mgter. en Informática y Computación
Profesor Adjunto
evalesani@exa.unne.edu.ar

Osvaldo P. Quintana
Experto en Estadística y Computación
Docente
oquin@mecon.indec.gov.ar

Oscar A. Vallejos
Mgter. en Informática y Computación
Profesor Adjunto
ovallejos@exa.unne.edu.ar

Universidad Nacional del Nordeste
Facultad de Ciencias Exactas y Naturales y Agrimensura
Dpto. de Informática
9 de Julio 1449
(3400) Corrientes, Argentina
(03783)-15-679884; (03783)-15-405117

Contexto

Miembros estudiantes de doctorado en Ingeniería de Software de la Universidad de Málaga (España) en convenio con las Universidades del Nordeste y Universidad de Misiones.

Resumen

El presente trabajo tiene por objeto la aplicación de Redes Neuronales Artificiales (RNA) como métodos de imputación, para ser utilizados sobre una base de datos real. Donde se simulo perdida de datos en distintos porcentajes, aplicando la técnica MCR (*Missing completely at random*). Estos datos faltantes o perdidos se completan mediante la aplicación de distintos modelos y en distintas situaciones, con el propósito de valorar el comportamiento de los mismos a través de distintos parámetros de eficiencias como MAE, MSE, y Regresión, se pretende determinar si RNA brinda una herramienta adecuada para la imputación en este caso en particular aplicados a datos de Censos Ganaderos.

Palabras clave: Imputación de datos, Redes neuronales artificiales, perceptrones multicapa, aprendizaje supervisado, imputación de datos en ganadería.

1. Imputación con redes Neuronales

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 [5]. 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 utilizan como por su amplia aplicabilidad, las redes perceptrón multicapa (MLP), consideradas aproximadores universales de funciones [4]. 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.

Generalmente los datos recopilados en censos y/o encuestas son procesados mediante técnicas estadísticas tradicionales. El inconveniente frecuente se presenta debido a la falta de respuesta por parte de los entrevistados. Se pensó que esta técnica de la IA presentaba una alternativa al momento de proporcionar imputación y/o predicciones sobre la existencia de ganado vacuno por medio de imputaciones anteriores en variables auxiliares, a partir de otros métodos de imputación, un archivo histórico u otros medios como opiniones de expertos temáticos.

Los datos empleados en el presente trabajo corresponden a los registrados para la provincia de Corrientes durante el Censo Nacional Agropecuario 2002.

En este trabajo, se describe el comportamiento de modelos de RNA supervisadas aplicadas en la imputación de Cabezas de Ganado Vacuno, partiendo como entrada a la red con los datos de Superficie dedicada a la Ganadería y el coeficiente de Hectáreas/Equivalente Vaca.

2. Líneas de investigación y desarrollo

La línea de investigación es imputación de datos con técnicas tradicionales [3] y no tradicionales como operadores de agregación de la mayoría [6] y RNA.

2.1. Descripción del experimento

Se describe la metodología aplicada en el diseño, desarrollo y entrenamiento de modelos de RNA supervisadas. A continuación se sintetizan las etapas consideradas:

Formulación del problema. La formulación del problema se concretó mediante la aplicación, en base a dos variables auxiliares, Superficie Dedicada a la Ganadería, y la relación de la superficie y el Equivalente Vaca Ha/EV, la incorporación de esta ultima variable auxiliar surgió como consecuencias de una entrevista con especialista de la temática específica, quien sugirió que actualmente se emplea como indicador importante el *equivalente vaca* que consiste en determinar la capacidad que tiene un campo para la ingesta de materia seca como alimentación del rodeo y la misma podría ser utilizada para predecir el número de ganado, de acuerdo a la zona donde se encuentre la explotación agropecuaria. Cabe aclarar que esta elección de variables evidénciales o variables de entrada se fundamenta en la no respuestas por diversos motivos de parte de los productores del sector a censos y encuestas especificas.

Selección de las variables evidenciales. Las variables relevantes requeridas para predecir el número de cabezas de ganado fueron obtenidas en base al conocimiento de

técnicos. Los valores que asumen las variables evidenciales corresponden a registros del Censo Nacional Agropecuario 2002.

Estudio para la implementación de una RNA: Se realizó un estudio referente a conceptos relacionados con la topología de una red neuronal, dimensionamiento de la red, arquitectura de una RNA, algoritmos de aprendizaje. Se seleccionó el aprendizaje supervisado por considerarse el más adecuado al problema y el que mayor validez tiene en amplios dominios del conocimiento.

Selección de las herramientas informáticas. Esta etapa se concretó mediante la elección de un toolbook para modelizar sistemas RNA. Estos toolboks o herramientas de software, son de carácter general, motivo por el cual deben evaluarse a fin de verificar que las opciones disponibles son adecuadas para el tratamiento de un problema en particular. Se examinaron las aplicaciones para la creación y tratamiento de modelos de redes neuronales artificiales disponibles. Los toolboks o herramientas existentes están destinadas a la predicción y/o clasificación. En la selección de las mismas se mencionan características como: la velocidad de ejecución, los requerimientos del sistema, las interfaces, el máximo número de variables y capas a especificar. Se seleccionó nntool de Matlab 7.4. Esta herramienta construye un modelo en base a los datos. Permite especificar diversos parámetros de aprendizaje. La conjunción de estas características se empleó para la definición de distintos modelos de entrenamiento de la RNA.

Construcción de modelos de RNA

a. **Arquitectura**. El formato del archivo para registrar los datos recolectados depende del modo de funcionamiento del software de RNA seleccionado. En primer lugar para crear los conjuntos de datos de entrenamiento y comprobación se diseñó el patrón (ejemplar o ejemplo). Cada patrón consta de dos partes. La primera parte es, un conjunto de números que representan los valores de las variables de entrada o evidenciales, empleadas para estimar los resultados. Si hay *m* variables predictoras, los primeros *n* elementos de cada patrón del archivo de datos de entrenamiento, segunda parte, la sección del “criterio" o “resultado", que consta de uno o más números, cada uno representa los valores de las variables de salida, es decir, los valores observados y que deberá predecir el modelo de RNA.

b. Definición de nodos de entrada y nodos de salida. Las variables evidenciales, SG = Hectáreas dedicada a la ganadería por cada explotación agropecuaria y Ha/EV = Hectáreas sobre Equivalente Vaca, se definieron como nodos de entrada y CG = el número de cabezas de ganado registrados en el Censo 2002 como valores esperados del nodo de salida.

c. **Diseño de los archivos de datos**. Se dividió el conjunto de datos en dos partes, una destinada al entrenamiento 75% y otros datos reservados para la comprobación de los modelos de RNA, 25%.

d. **Entrenamiento de los modelos.** Consiste en la definición de la topología de la RNA. No existe un procedimiento específico para definir *a priori* del número de capas ocultas y el número de neuronas por capa necesarios para lograr el aprendizaje de la RNA. Se propusieron distintas configuraciones modificando el número de nodos y capas intermedias, funciones de activación y parámetros.

e. **Ejecución de los modelos**, se entrenaron distintos modelos de RNA utilizando el software nntool de Matlab 7.4.

f. **Validación de los modelos**. En el aprendizaje supervisado, una medida de calidad del modelo está dada en términos de los valores del error estándar MSE y el MAE [2] y la regresión entre salida de la red y salida real que proporciona la herramienta utilizada para el modelado de la red.

Validación del software. Finalizado el desarrollo de los modelos, es imprescindible verificar el correcto funcionamiento del mismo. Se deben implementar validaciones internas y validaciones externas. Las primeras fueron realizadas por los autores del trabajo. Las segundas se llevarán a cabo con especialistas en la temática.

3. Resultados obtenidos

A partir de los datos proporcionados por el Censo Nacional Agropecuario 2002, se buscó modelizar RNA para imputar la cantidad de cabezas de ganado de una explotación a partir de los datos de Superficie Total de la explotación Agropecuaria, Superficie Dedicada a la Ganadería y Estrato de tamaño entre otras variables disponibles, se descartaron muchos modelos entrenados, por considerar que los resultados obtenidos no fueron aceptables en función a los valores de los parámetros utilizados para medir su eficiencia como el error cuadrático medio, el error absoluto y regresión.

Posteriormente y ante la consulta con expertos en el tema se considero finalmente como variable evidenciales a la Hectáreas Dedicada a la Ganadería (SG) y el Coeficiente Hectárea sobre Equivalente Vaca (Ha/EV), consiguiendo con este modelo una RNA que brindo restados muy aceptables para la imputación de Cabezas de Ganados (CG).

Entrada

Superficie dedicada a la ganadería = SG
Hectárea por Equivalente Vaca = Ha/EV

Salida

Cantidad de Cabezas de Ganado = CG

Se efectúo en principio un análisis exploratorio de los datos mediante el BoxPot y se descartaron los datos fuera del rango para el ítems relación Ha/EV donde se separo alrededor de 100 establecimiento, que representaba el 0,22 % del total de cabezas, Se verifico si existía algún de relación entre las variables de entradas y se observo que no estaban correlacionadas ya que el coeficiente de Pearson tenia el valor de 0.078, prácticamente igual a cero, se realizo la misma comprobación entre la variable de entradas SG, Ha/EV y la salida CG, encontrando que la SG tenia un relación alta que rondaba el valor de 0.90 de mismo coeficiente, mientras que la variable de salida con la relación Ha./EV tenia una relación baja.

Se experimentó con redes Multicapas con distintas estructuras de capas, con 1,2 y 3 capas intermedias, distintas funciones de transferencias y aprendizajes supervisados de retropropagación, con funciones varias funciones de aprendizaje, con datos reales sin transformaciones y posteriormente con la normalización de los datos con sus mínimos y máximos, hasta encontrar una red que con tres capas intermedia de topología 2-14-8-4-1 con función de aprendizaje LEVENBERG –MAQUARD y la función Logística como función de transferencias para todas las capas, la cuál brindo resultados satisfactorios para la etapa de entrenamiento con el 75% de los datos consiguiéndose una buena convergencia del error como puede observarse en la Figura 1, aplicando la misma al 25% restante que se reservó para realizar el test de comprobación, al cual se aplico una medida de regresión entre la salida real y la obtenida por la RNA , consiguiendo un valor de R= 0.982, que se considera muy aceptable en términos estadísticos, como se observa en la Figura 2.

4. Conclusiones

La imputación mediante RNA se diferencia fundamentalmente lo métodos estadísticos tradicionales, en que los primeros no realizan condicionamiento, ni ninguna hipótesis sobre

la distribución de los datos a estudiar y que en este caso en particular sobre la imputación de Cabezas de Ganado Vacuno brinda una herramienta que merece ser tenida en cuenta para el tratamiento de datos faltante como método de imputación en Encuestas y Censos del sector, principalmente si se cuenta con las variable auxiliares necesarias para su aplicación, de las cuales van a depender directamente la imputación a realizar, las que se pueden obtener por

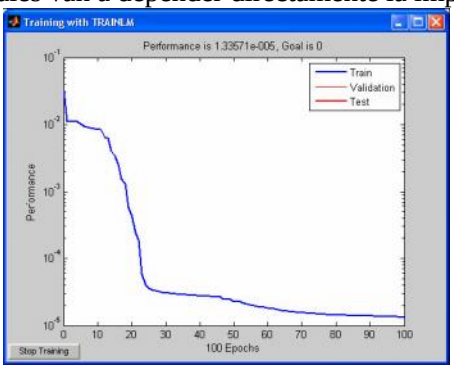

Figura 1: Error de Entrenamiento de la Red (100 Epochs)

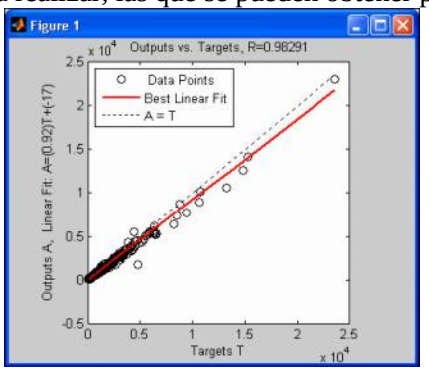

Figura 2: Comparación salida de la RNA y salida Real

medio de trabajos realizados con anterioridad, por consultas con expertos, mediante estudio de cartografía digitales con fotointerpretación de las explotaciones, material del cual se dispone en la provincia donde se realizo el estudio o por imputación con otros métodos como imputación multitarea con redes neuronales [1] siendo la misma una línea de futuros trabajos, con el fin de poder contar con datos más precisos sobre el tema tratado en este caso, ya que representa de sumo interés en la actualidad debido a la preponderancias que tienen los productos de origen primarios en el sistema económico nacional y global.

5. Referencias

[1] Caruana R. (1993), Multitask learning: a knowledge-based source of inductive bias. Proceedings of the 10th International Conference of Cognitive Science, pp. 41-48.

[2] Castillo, E., Cobo, A.; Gutiérrez, J. M. y Pruneda, R. E. (1999). “Introducción a las redes funcionales con aplicaciones. Un nuevo paradigma funcional”. Ed. Paraninfo.

[3] Doña J.M., Quintana O.P., Valesani M.E., Vallejos O.A. (2008) “Analysis of Agregation Methods in Incomplete Database Systems”. Information Processing and Management of Uncertainty in Knowledge-Based System (IPMU 2008). ISBN: 978-84-612-3061-7

[4] 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.

[5] 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).

[6] Vallejos, Oscar A, Valesani, Maria E, Quintana, Osvaldo (2008) “Imputación de datos con operadores OWA de la mayoría”. X Workshop de Investigación en Ciencias de la Computación Gral Pico La Pampa Argentina 5 Y 6 de Mayo 2008 Org Universidad Nacional de La Pampa - Facultad de Ingeniería.