Segmentación de Imágenes Mediante Redes Neuronales para el Análisis de Señales Acústicas Emitidas por Cetáceos

Rosa María Menchón Lara, María Consuelo Bastida Jumilla, Juan Morales Sánchez, José Luis Sancho Gómez.
Departamento de Tecnologías de la Información y las Comunicaciones
Universidad Politécnica de Cartagena. Plaza del Hospital, Nº 1, 30202 Cartagena (Murcia)
Teléfono: 968326542
E-mail: rmml@alu.upct.es

Resumen. *Las técnicas de monitorización acústica de mamíferos marinos han adquirido una gran relevancia. Este trabajo se centra en la detección de las señales tonales (silbidos y sonidos modulados de baja frecuencia) emitidas por cetáceos, que componen el orden de mamíferos marinos más vocalizantes. Se propone un nuevo método de segmentación de las vocalizaciones de un espectrograma basado en la utilización de redes neuronales. En particular, un comité de redes neuronales es empleado para este fin, entrenando todas las redes mediante el algoritmo OP-ELM. Una vez completado el proceso de entrenamiento, el sistema propuesto permite obtener las imágenes segmentadas de forma automática.*

1 Introducción

La monitorización acústica de mamíferos marinos es de gran importancia para la detección, localización e identificación de estas especies [5]. Permite recopilar información para el estudio de los hábitos y las habilidades de comunicación de los mamíferos marinos. Además, tener constancia de la presencia de estos animales en ciertas áreas facilita su protección.

Las señales acústicas emitidas por los mamíferos marinos son denominadas *vocalizaciones* y se clasifican como *señales tonales* (silbidos y sonidos modulados de baja frecuencia) o *señales pulsadas* (chasquidos). Los sonidos pulsados son señales de banda ancha y de muy corta duración; mientras que los sonidos tonales son señales de banda estrecha moduladas en frecuencia, en las que frecuencia y amplitud varían lentamente con el tiempo y el rango de duración es mucho más amplio. Este trabajo se centra en las señales tonales emitidas por cetáceos, que componen el orden de mamíferos marinos más vocalizantes. La mayoría de los sistemas de detección de mamíferos marinos se basan en la utilización del *espectrograma*, ya que se trata de una transformación tiempo-frecuencia con una sólida base matemática y resulta fácilmente interpretable.

En este artículo, se propone un nuevo método de segmentación de las vocalizaciones presentes en un espectrograma basado en la utilización de redes neuronales. Concretamente, se ha empleado una combinación de *Perceptrones Multicapa*, que han sido entrenados mediante el algoritmo *Optimally Pruned Extreme Learning Machine* [2].

En nuestro caso, la tarea de segmentación es considerada como un problema de reconocimiento de patrones. El sistema asigna un valor a cada píxel de la imagen de entrada (espectrograma), indicando la presencia de vocalizaciones de cetáceos.

2 Multilayer Perceptrons

Una importante categoría de redes neuronales de tipo "*feed-forward*" son los Perceptrones Multicapa (*MultiLayer Perceptron*, MLP). Los MLPs han sido aplicados satisfactoriamente para resolver multitud de problemas, entrenándolos de forma supervisada. La arquitectura de un MLP está completamente definida por una capa de entrada, una o más capas ocultas y una capa de salida. Estando cada capa compuesta por, al menos, una neurona. La señal de entrada se propaga capa por capa hacia la salida de la red [4].

En este trabajo, vamos a considerar redes neuronales de una única capa oculta (*Single Layer Feedforward Networks*, SLFNs). Supongamos un MLP con una capa oculta de M neuronas. Cada neurona de la capa oculta es conectada a todas las neuronas de la capa anterior y posterior a ella, véase la Fig. 1.

Dado un conjunto de datos compuesto por N muestras $(\boldsymbol{x}_i, \boldsymbol{t}_i)$, donde $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \dots, x_{in}]^T \in \mathbb{R}^n$ (vectores de entrada) y $\boldsymbol{t}_i = [t_{i1}, t_{i2}, \dots, t_{im}]^T \in \mathbb{R}^m$ (vectores "*target*", salidas deseadas), la salida del MLP viene dada por

$$\boldsymbol{o}_i = \sum_{j=1}^{M} \boldsymbol{\beta}_j f(\boldsymbol{w}_j \cdot \boldsymbol{x}_i + b_j), \qquad (1)$$

donde $f(\cdot)$ son las funciones de activación de las neuronas ocultas, $\boldsymbol{\beta}_j = [\beta_{j1}, \beta_{j2}, \dots, \beta_{jm}]^T$ son los pesos de salida asociados a la neurona oculta j-ésima; mientras que $\boldsymbol{w}_j = [w_{j1}, w_{j2}, \dots, w_{jn}]^T$ y b_j representan, respectivamente, el vector de pesos de entrada y el sesgo correspondientes a dicha neurona. Las neuronas de salida tienen funciones de activación lineales, mientras que para $f(\cdot)$ se suelen emplear funciones de tipo sigmoide.

Los parámetros a optimizar durante el entrenamiento de la red son los pesos y el número de neuronas de la

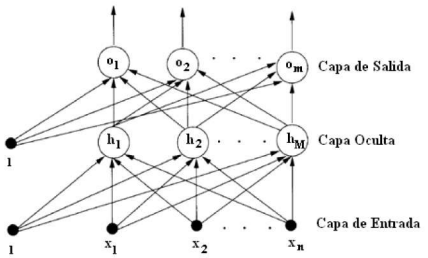

Fig. 1. Arquitectura de un MLP con una única capa oculta.

capa oculta. Para más información sobre el diseño y la implementación de MLPs ver [4].

3 Extreme Learning Machine

En general, una SLFN estándar con M neuronas ocultas y función de activación $f(\cdot)$ puede aprender de forma exacta N patrones distintos (con $N \geq M$), inicializando aleatoriamente los pesos de entrada y el sesgo de las neuronas ocultas. Esta asignación aleatoria se puede llevar a cabo si f es infinitamente diferenciable [1].

El algoritmo *Extreme Learning Machine* (ELM) se fundamenta en este hecho. Tras inicializar de manera aleatoria los pesos de entrada, considera la SLFN como un sistema lineal y los pesos de salida pueden obtenerse de forma sencilla. La solución analítica del problema se calcula a través de la pseudo-inversa de la matriz de salidas de la capa oculta ($\boldsymbol{H} \in \mathbb{R}^{N \times M}$).

De esta forma, dado un conjunto de N vectores de entrada, la salida de la SLFN aproxima estos $\mathbf{N}$ casos con error nulo ($\sum_{i=1}^{N} \|\boldsymbol{y}_i - \boldsymbol{t}_i\| = 0$), es decir, existen $\boldsymbol{\beta}_j$, $\boldsymbol{w}_j$ y b_j tal que,

$$\sum_{j=1}^{M} \boldsymbol{\beta}_j f(\boldsymbol{w}_j \cdot \boldsymbol{x}_i + b_j) = \boldsymbol{t}_i, \quad i = 1, \dots, N. \tag{2}$$

Las anteriores N ecuaciones se pueden expresar de forma compacta de la siguiente forma:

$$\boldsymbol{H\beta} = \boldsymbol{T}, \tag{3}$$

donde $\boldsymbol{\beta} \in \mathbb{R}^{M \times m}$ es la matriz de pesos de salida, y $\boldsymbol{T} \in \mathbb{R}^{N \times m}$ es la matriz de targets. Según esto, el entrenamiento de la red se corresponde con la solución del problema de mínimos cuadrados establecido en (3). Así, los pesos óptimos de la capa de salida son $\widehat{\boldsymbol{\beta}} = \boldsymbol{H}^{\dagger}\boldsymbol{T}$, donde $\boldsymbol{H}^{\dagger}$ es la pseudo-inversa de Moore-Penrose.

El ELM es fácil de utilizar y proporciona un entrenamiento rápido y eficiente. Sin embargo, es necesario fijar la arquitectura de la red, es decir, el número de neuronas ocultas. En general, se desconoce el valor óptimo de M y es necesario realizar una búsqueda del mismo mediante validación cruzada, lo que resulta costoso computacionalmente.

4 Optimally Pruned ELM

El método *Optimally Pruned ELM* (OP-ELM) [2] selecciona el tamaño óptimo de la red de forma automática. Inicialmente, fija un número de neuronas ocultas muy elevado y construye la red mediante el algoritmo ELM estándar. A continuación, es aplicado el algoritmo *Least Angle Regression* (LARS) a fin de obtener un ranking de las neuronas según su relevancia para resolver el problema de mínimos cuadrados. La solución obtenida por LARS es única cuando el problema es lineal [3]. Puesto que, en nuestro caso, la salida es lineal con respecto a las neuronas ocultas, el orden de neuronas proporcionado por LARS será exacto.

Una vez ordenadas, se eliminan las neuronas menos útiles mediante validación cruzada del tipo *Leave One Out* (LOO), escogiendo aquellas que proporcionan un menor error de validación. Por tanto, sólo son seleccionadas las M^* neuronas más importantes según LARS (con $M^* < M$), y se obtiene una solución única para el diseño de la red neuronal.

5 Comités de Redes Neuronales

En la práctica, los problemas de clasificación complejos requieren la contribución de varias redes neuronales para alcanzar una solución óptima [4]. De acuerdo con el principio "*divide y vencerás*", una tarea compleja se resuelve dividiendo ésta en varias tareas más sencillas y combinando sus soluciones. En la terminología de las redes neuronales, esto se traduce en dividir la tarea de aprendizaje entre varios expertos. Esta combinación de expertos se denomina comité de redes neuronales y se incluye en la categoría de aproximadores universales.

En el campo del procesado de imagen, se ha demostrado que la precisión de clasificación proporcionada por un comité puede superar la precisión individual de la mejor red. Sin embargo, también se ha demostrado que sólo son efectivos si las redes que los forman producen errores diferentes. Se han investigado diferentes métodos para la combinación de redes neuronales. Éstos consisten, básicamente, en variar los parámetros relacionados con el diseño y la implementación de los expertos.

6 Método Propuesto

El objetivo principal de este trabajo es extraer de un espectrograma las señales tonales emitidas por los cetáceos. Con este propósito, se propone un método de segmentación basado en el entrenamiento de redes neuronales. Dada una imagen de entrada, el sistema debe clasificar sus píxeles en dos clases: por un lado los pertenecientes a las vocalizaciones, y por otro el resto de píxeles, que no contienen información de interés. Con el entrenamiento apropiado, el sistema es capaz de detectar automáticamente las señales tonales de un espectrograma.

La segmentación no es una tarea trivial y, además, debemos tener en cuenta las características de las imágenes con las que trabajamos. En los espectrogramas, las señales se encuentran inmersas en un entrono de ruido enmascarante (ruido de fondo y reverberación) que complica aún más el problema. A fin de mejorar la precisión y robustez del sistema, los resultados de cuatro redes diferentes han sido

combinados, es decir, se ha utilizado un comité de redes (ver Fig. 2).

Las redes neuronales utilizadas son MLPs con una única capa oculta. Las redes han sido entrenadas mediante el algoritmo OP-ELM y, por tanto, el tamaño óptimo de la capa oculta es automáticamente seleccionado. El sistema utiliza el espectrograma como imagen de entrada y una imagen etiquetada manualmente como salida deseada (imagen "target"). La imagen target es una imagen binaria en la que los píxeles con valor '1' muestran las señales tonales emitidas por los cetáceos (ver Fig. 3).

Los cuatro expertos de nuestro comité (de RN 1 a RN 4 en la Fig. 2) toman como entradas sub-imágenes resultantes de un *proceso de enventanado*. Una ventana cuadrada es desplazada píxel a píxel sobre la imagen original para construir los vectores de entrada a la red. Para cada sub-imagen de entrada, la salida de la red presenta una única componente, que se corresponde con el valor target para el píxel central. Diferentes tamaños de ventana (W = 5, 7, 9 y 11) han sido considerados para construir el conjunto de entrenamiento de los cuatro expertos.

Por último, una "*meta*" red neuronal combina los resultados obtenidos por los expertos. Esta red también es entrenada con OP-ELM, pero en una segunda etapa, pues son necesarios los resultados de los expertos. El vector de entrada asociado a un píxel se construye a partir de cuatro sub-imágenes de tamaño 9x9, una de cada experto. La red proporciona como salida el valor target asociado al píxel bajo análisis, es decir, una única componente de salida.

7 Resultados

Una vez completado el proceso de entrenamiento de todas las redes de nuestro comité, realizamos las simulaciones pertinentes para comprobar la precisión del método propuesto. La Fig. 3 muestra la imagen de entrada (espectrograma) y la imagen target. La imagen segmentada obtenida a la salida del comité de redes se presenta en la Fig. 4. Como se puede apreciar, estos resultados, aunque preliminares, son bastante satisfactorios.

8 Conclusiones

En este artículo se propone un método de segmentación basado en redes neuronales aplicado a la detección de las señales tonales emitidas por cetáceos. Una combinación de MLPs ha sido entrenada mediante el algoritmo OP-ELM para llevar a cabo dicha tarea, constituyendo lo que se conoce como un comité de redes neuronales.

Una vez completado el entrenamiento, el sistema propuesto permite obtener la segmentación de los espectrogramas de forma automática. Además, el proceso de aprendizaje de las redes presenta los beneficios propios del algoritmo OP-ELM: fiabilidad, rapidez y diseño óptimo de la red.

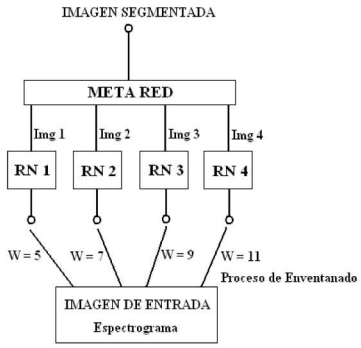

Fig. 2. Esquema del comité de redes propuesto.

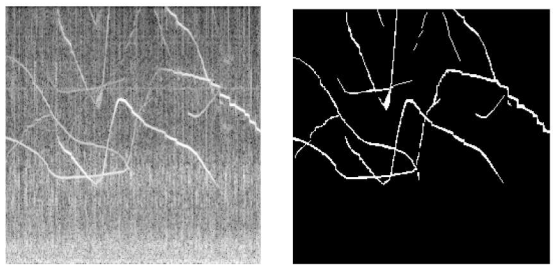

Fig. 3. Imagen de entrada (izquierda) e imagen target (derecha).

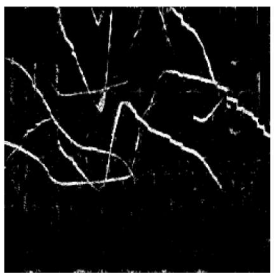

Fig. 4. Imagen segmentada, salida del sistema.

Los resultados preliminares obtenidos son muy prometedores y es necesario continuar trabajando en un proceso de validación de los mismos para verificar la robustez y capacidad de generalización del método propuesto.

Agradecimientos

Este trabajo está parcialmente financiado por el MEC a través del proyecto TEC2009-12675, por la UPCT (Iniciación a la Actividad Investigadora) y por la Fundación Séneca (09505/FPI/08).

Referencias

[1] G. B. Huang et al., "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, 489-501, 2006.

[2] Y. Miche et al., "OP-ELM: Optimally pruned extreme learning machine," *IEEE Transactions on Neural Networks*, vol. 21, 158-162, 2010.

[3] B. Efron et al., "Least angle regression," *Annals of statistics*, vol. 32, 401-499, 2004.

[4] S. Haykin, *Neural Networks: A comprehensive foundation*. Prentice-Hall, 1999.

[5] A. Sánchez, P. Muñoz, J. L. Sancho, "A novel image-processing based method for the automatic detection, extraction and characterization of marine mammal tonal calls," *Journal of the Marine Biological Association of the United Kingdom*, vol. 90, 1667-1684, 2010.