Análisis ROC Aplicado al Modelo de Computación Blanda para Predecir la Percepción del Habla en Niños con Implante Coclear.

Trabajo de Tesis
presentado al
Departamento de Ingeniería Industrial

por

Liliana Eugenia Calderón Jerez

Asesor: Andrés Medaglia PhD.
Co-asesora: Olga Lucía Sarmiento PhD.

Para optar al título de
Ingeniera Industrial

Ingeniería Industrial
Universidad de los Andes
Mayo 2006.

Agradecimientos

Agradezco a todas las personas que directa o indirectamente, estuvieron involucradas en el desarrollo de este proyecto. Especialmente al profesor Andrés Medaglia, a Olga Lucía Sarmiento, investigadora principal del proyecto Implante Coclear, a Isabel Cristina Ramírez y al equipo de fonoaudiólogos de la Fundación Santa Fe de Bogotá, por compartirme todo su conocimiento y experiencia. Me siento muy satisfecha de haber podido, a través de mi trabajo, beneficiar a niños que por algún motivo sufren de sordera profunda.

Adicionalmente, quiero agradecerle a mi familia por el respaldo que siempre me han brindado, sin el cual, no sería ni la persona, ni la profesional que hoy en día soy.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Ecuaciones

Capítulo I

Introducción

Este trabajo de gado, es uno de un grupo de trabajos que se han realizado para el Proyecto Implante Coclear, el cual ha sido desarrollado por las Facultades de Medicina e Ingeniería Industrial de la Universidad de los Andes. Dicho Proyecto, busca proponer un modelo que prediga la percepción del habla en niños con sordera profunda, una vez obtengan el implante coclear.

Jorge Rodríguez D'Alleman [1] realizó uno de estos trabajos, en el cual implementó un modelo de pronóstico clínico combinando Redes Neuronales Artificiales (RNA) y Algoritmos Genéticos (AG), aplicado a variables que miden la percepción del habla en niños con el implante. Las variables de respuesta, escogidas a criterio médico, fueron medidas a los 24 meses del implante y son: Discriminación de Bisílabos, Discriminación de Frases y Promedio Tonal. De esta manera, se busca pronosticar el comportamiento del paciente para cada una de las anteriores variables de respuesta, a través del modelo propuesto, dadas ciertas variables de entrada. El criterio usado para optimizar la configuración de las redes neuronales, fue minimizar la raíz del error cuadrático medio (RECM), la cual es una medida que compara la predicción del modelo para cada una de las variables de respuesta, contra el comportamiento real de los pacientes en la misma variable de respuesta.

El objetivo de este proyecto de grado es evaluar el desempeño del modelo de pronóstico clínico propuesto en el proyecto de Rodríguez D'Alleman, a través del análisis ROC (Receiver Operating Characteristic). Este análisis se basa en la construcción de la curva ROC para los modelos de predicción y la estimación del área bajo la misma. La utilidad de ésta aproximación radica, en la posibilidad de medir la precisión predictiva del modelo a través de una sola cifra. El área bajo la curva puede ser un valor entre 0.5 y 1, donde 0.5 indica que la prueba no tiene ningún poder predictivo y 1 que la prueba discrimina de manera perfecta los pacientes.

Un enfoque tradicional de pronóstico en aplicaciones clínicas, es la Regresión Logística Multivariada, la cual es un método de pronóstico por sí misma. En este caso, se corren todas las variables de entrada en conjunto para lograr un modelo de predicción [1]. Se usará el análisis ROC para comparar el desempeño del modelo tradicional con el de la red neuronal.

Con la intención de guiar al lector, el documento se desarrolla de la siguiente manera: en el Capítulo II se presenta el marco teórico del análisis ROC, el Capítulo III cubre la implementación del algoritmo propuesto, en el Capítulo IV se exponen los resultados obtenidos y se comparan con los resultados de la regresión logística, en el Capítulo V se encuentran las conclusiones del análisis junto con la investigación futura que se puede generar a partir de este proyecto de grado.

Capítulo II

Marco Teórico

2.1. Análisis ROC

Resulta interesante el surgimiento del análisis ROC, ya que la motivación inicial para su desarrollo fue muy diferente a su actual aplicación. En los años 40, luego del ataque japonés a la base naval de Pearl Harbor, Estados Unidos quiso determinar la razón por la cual sus radares "Receiver Operators" (de allí se deriva su nombre) no detectaron el acercamiento de la flota japonesa. De esta manera, las curvas ROC hicieron parte del desarrollo de sistemas discriminatorios que detectarán señales de radio en presencia de ruido. A partir de los años 60, el análisis ROC empezó a ser usado en investigaciones médicas y se ha convertido en una herramienta de gran utilidad para los médicos en la toma de decisiones.

La curva ROC es un método útil para evaluar el desempeño de una prueba diagnóstica. Esta estadística se define como la representación gráfica de la relación entre sensibilidad (tasa de predicción de verdaderos positivos) y especificidad (tasa de predicción de verdaderos negativos). De esta manera, la sensibilidad indica qué proporción de los pacientes que realmente tendrán una respuesta positiva, serán detectados por la prueba. Y la especificidad, indica qué proporción de los pacientes que tendrán una respuesta negativa, serán detectados por la misma prueba.

Como se observa en la Figura 1, un punto sobre la curva ROC corresponde a una coordenada de 1-especificidad y sensibilidad, la primera ubicada en el eje de X y la segunda en el eje de Y. Cada punto (1-especificidad, sensibilidad) sobre la curva se obtiene variando el punto de corte que determina si el resultado de la prueba será considerado como positivo o negativo.

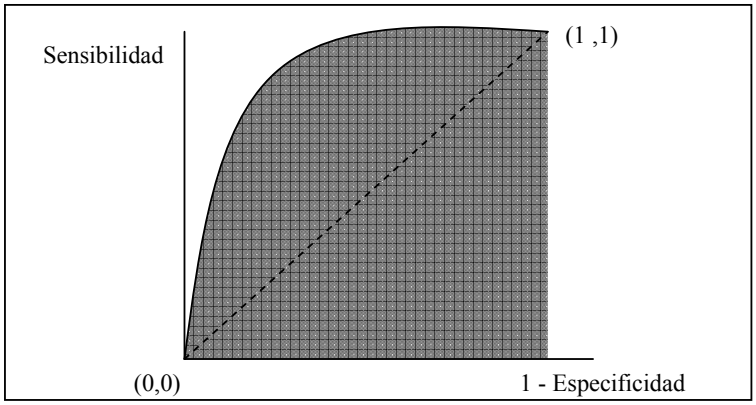


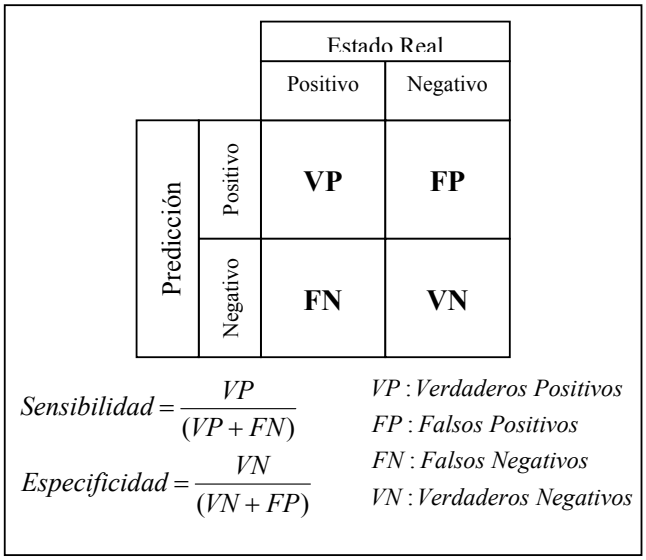
Figura 1. Representación gráfica de la Curva ROC

Por lo tanto, es necesario fijar un conjunto de puntos de corte que pertenezcan al rango de resultados de la prueba diagnóstica, y calcular para cada uno de ellos, la correspondiente sensibilidad y especificidad. Estas dos medidas se hallan contrastando el verdadero estado de cada paciente (positivo o negativo), con el resultado (positivo o negativo) de la prueba diagnóstica. El criterio para determinar si el estado real y los resultados de la prueba son positivos o negativos, es el punto de corte. De esta manera, si el resultado es menor al punto de corte, éste se considera como negativo -0- y si es mayor o igual se considera positivo -1- (aunque en algunas ocasiones, puede ser de manera inversa). Por lo tanto, para cada punto de corte escogido, se obtendrá un punto sobre la curva ROC.

Junto a la Ecuación 1 se muestra la tabla de 2x2 que clasifica cada una de las observaciones (pacientes), de acuerdo a su estado real en la variable de respuesta y la predicción obtenida con el modelo correspondiente. Al contrastar los datos para todas las observaciones, se obtienen los valores de VP, FP, FN y VN, con los cuales es posible calcular la tasa de predicción de verdaderos positivos (sensibilidad) y verdaderos negativos (especificidad).

Una vez se han graficado todas las coordenadas (1-especificidad, sensibilidad), se aplica la regla trapezoidal para estimar el área bajo la curva, la cual consiste en sumar las áreas entre cada par de puntos sobre la curva ROC. El área bajo la curva ROC se interpreta como la probabilidad de identificar correctamente un paciente que presentará una respuesta positiva

de uno que tendrá una negativa. Esta estimación es un valor entre 0.5 y 1, donde 0.5 indica que la prueba no tiene ningún poder predictivo, y la curva ROC quedaría representada como la línea recta de 45° que se muestra en la Figura 1.



Ecuación 1. Cálculo de Sensibilidad y Especificidad

Un área igual o muy cercana a 1, indica que la prueba siempre pronostica acertadamente el comportamiento de los pacientes, posterior al implante coclear.

2.2. Comparación Estadística entre Curvas ROC

La comparación del desempeño de la Regresión Logística Multivariada, con el desempeño de la Red Neuronal, se realiza estimando el Error Estándar (SE) del área bajo la curva ROC de ambos modelos de predicción. Con cada uno de los SE, es posible comparar estadísticamente la diferencia de medias de ambas áreas. El error estándar se estima de la siguiente manera [3]:

$$SE = \sqrt{\frac{A(A-1) + (N_p - 1)(Q_1 - A^2) + (N_n - 1)(Q_2 - A^2)}{N_p N_n}}$$

$$Q_1 = \frac{A}{(2-A)} \qquad Q_2 = \frac{2A^2}{(1+A)}$$

A: área bajo la curva
N_p: Número de observaciones positivas (≥ Punto de Corte)
N_n: Número de observaciones negativas (<Punto de Corte)

Ecuación 2. Estimación del Error Estándar

Cuando ambas pruebas diagnósticas se aplican al mismo conjunto de pacientes, es necesario tener en cuenta la correlación que los mismos pacientes incluyen al modelo. De esta manera, el SE para la diferencia de áreas se estima así [4]:

$$SE(A1 - A2) = \sqrt{SE(A_1)^2 + SE(A_2)^2 - 2r * SE(A_1)SE(A_2)}$$

SE (A_1): Error Estándar del área 1 (Red Neuronal).
SE (A_2): Error Estándar del área 2 (Regresión Logística).
r: correlación inducida por estudiar el mismo conjunto de pacientes.

Ecuación 3. Estimación del Error Estándar para la diferencia de áreas

Para hallar el valor de la correlación r, es necesario hallar primero los coeficientes de correlación r_p y r_n. El primer coeficiente corresponde a la correlación entre los pacientes positivos (1) y el segundo a la correlación entre de los pacientes negativos (0). Para calcular estos dos coeficientes, se usa la aproximación Pearson product-moment debido a que la curva ROC se deriva de datos clasificados por intervalos. En este caso, la regresión logística y la red neuronal, usan los mismos datos de los mismos pacientes. Por lo tanto, la aproximación de Pearson se reduce a la fórmula mostrada en la Ecuación 4.

$$r = \frac{\sum (X - \mu_x)(Y - \mu_y)}{N\sigma_x\sigma_y} = \frac{\sum (X - \mu_x)^2}{N\sigma^2{}_x}$$

r : coeficiente de correlación

X : Resultados reales de los pacientes usados para la prueba X

Y : Resultados reales de los pacientes usados para la prueba Y

N: Número de observaciones

μ_x : Media de los resultados reales usados para la prueba X

μ_y : Media de los resultados reales usados para la prueba Y

σ_x : Desviación Estándar de los datos usados para la prueba X

σ_y : Desviación Estándar de los datos usados para la prueba Y

Ecuación 4. Aproximación Pearson product-moment.

Finalmente, se calcula el promedio de r_n y r_p, así como el promedio de las áreas para los dos modelos predictivos. En la tabla de coeficientes de correlación presentada por Hanley y McNeil [4], es posible Hallar el valor de r.

El estadístico de prueba que determina si las áreas son estadísticamente iguales o diferentes se muestra en la Ecuación 5.

$$Z = \frac{(A_1 - A_2)}{SE(A_1 - A_2)}$$

Ecuación 5. Estadístico de prueba para la diferencia de áreas

Capítulo III

Implementación

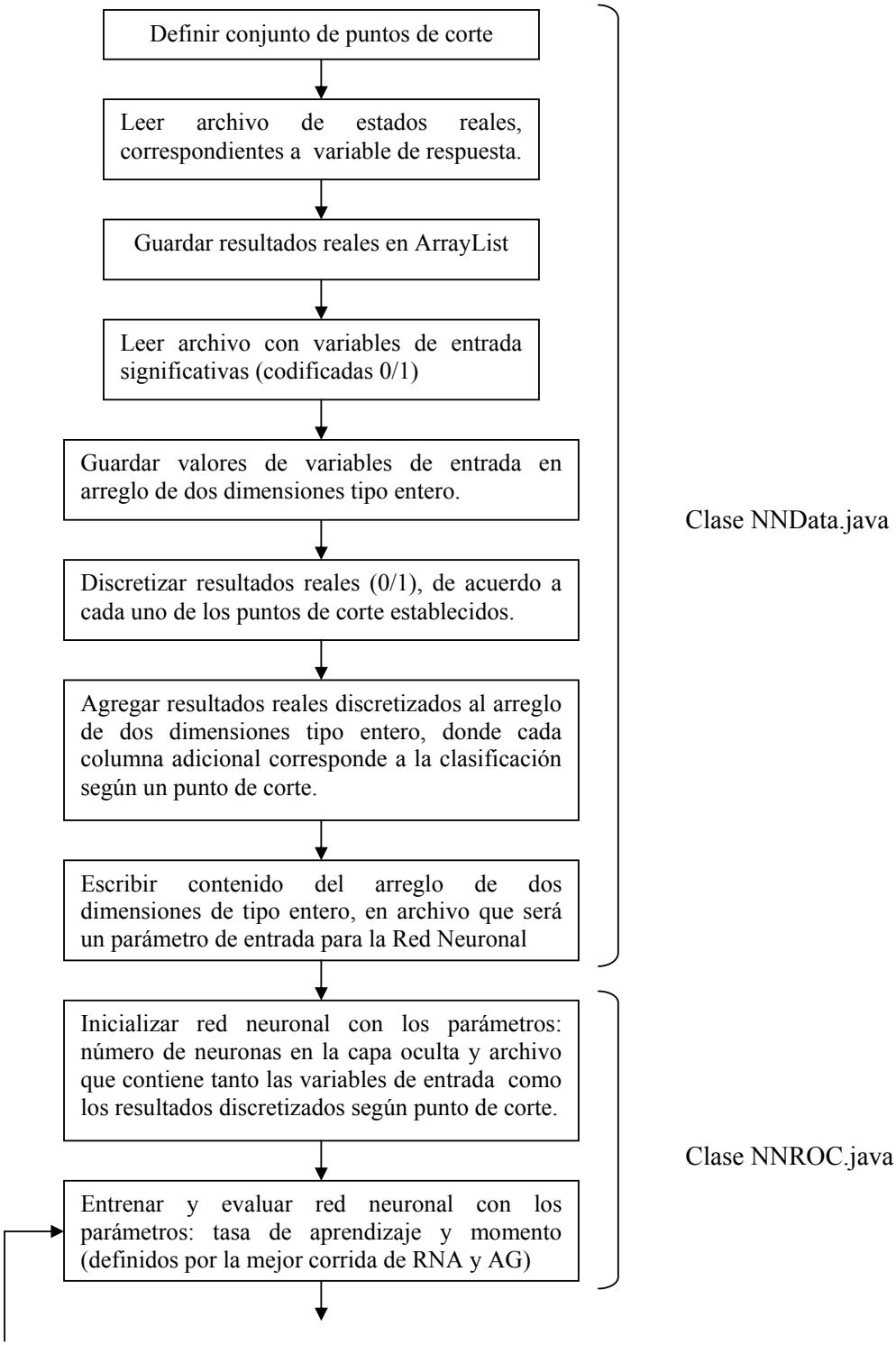
Los valores de los pronósticos realizados por las RNA para las tres variables de respuesta son continuos y se encuentran en los siguientes rangos: discriminación de bisílabos y discriminación de frases en contexto abierto sen encuentran entre 0 y 100%; el rango de la variable promedio tonal se encuentra entre 0 y 50 decibeles. Cada variable de respuesta es codificada de manera binaria (0/1), usando un valor crítico o punto de corte. En el caso de las dos primeras variables de respuesta, cuando el pronóstico del modelo es mayor o igual al punto de corte, la respuesta se considera positiva (1), de lo contrario se considera negativa (0). Por otro lado, la dicotomización de la variable promedio tonal se determina de manera contraria, cuando el pronóstico es menor o igual al punto de corte la respuesta se toma como positiva (1) y cuando es mayor se toma como negativa (0), tal y como se muestra en la Tabla 1.

Variable de respuesta	Pronóstico Variable < Punto de Corte	Pronóstico Variable >= Punto de Corte
Discriminación de Bisílabos	Negativo (=0)	Positivo (=1)
Discriminación de Frases	Negativo (=0)	Positivo (=1)
Promedio Tonal	Positivo (=1)	Negativo (=0)

Tabla 1. Dicotomización de las Variables de Respuesta

En la Figura 2 se muestra el diagrama de flujo con el macro algoritmo propuesto para la programación del análisis ROC. El código correspondiente a este algoritmo, fue escrito en lenguaje Java e implementado en el editor Eclipse SDK. Se eligió este lenguaje por su facilidad para escribir, leer, invocar objetos ya creados y especialmente por su portabilidad (“Program once, Run anywhere”). Los resultados correspondientes a la regresión logística, se obtuvieron usando el paquete estadístico SAS.

3.1. Macro Algoritmo



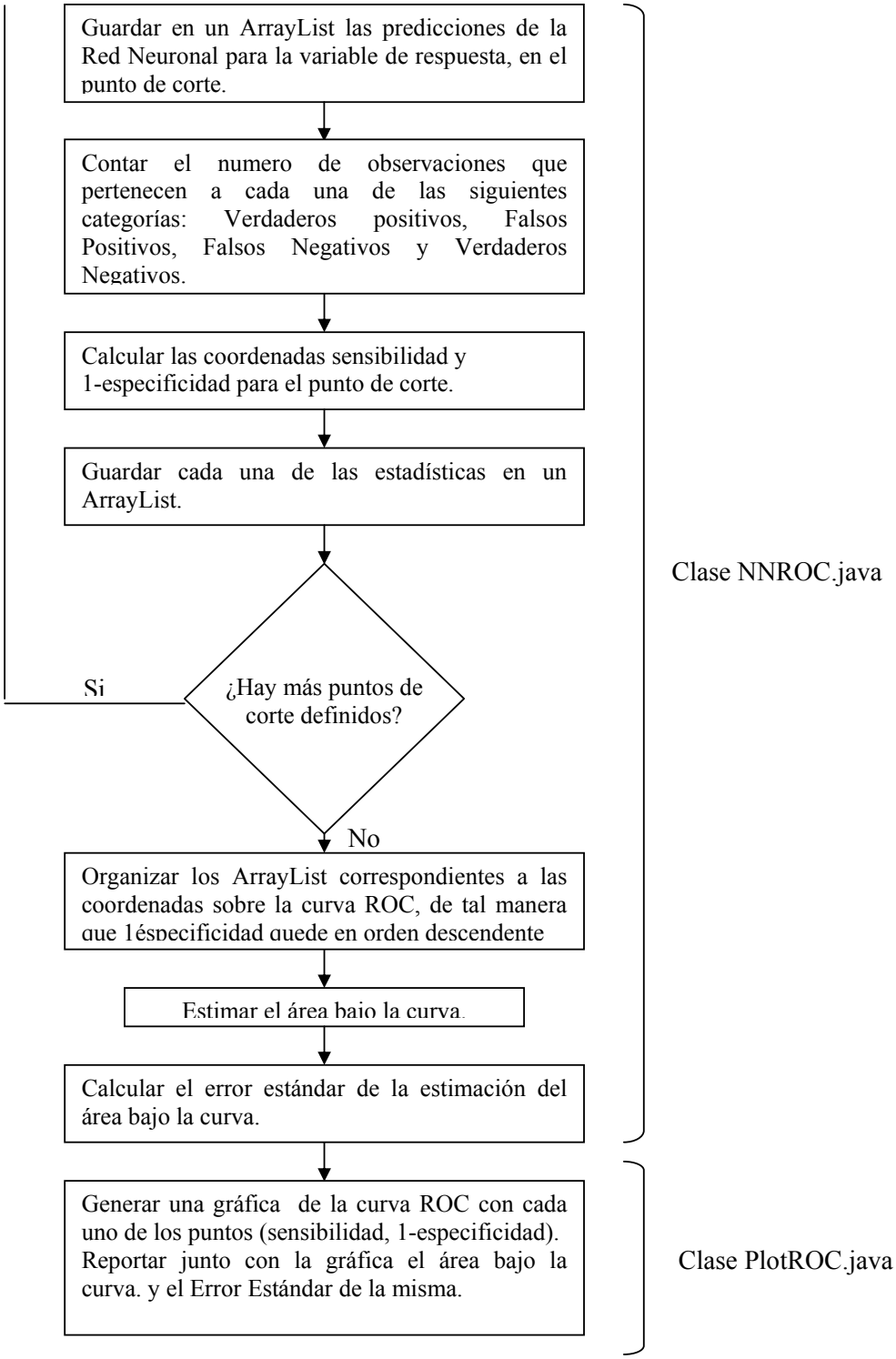


Figura 2. Diagrama de flujo con macro algoritmo propuesto.

Para cada uno de los pasos del anterior algoritmo, se crearon métodos agrupados de tal manera que tareas similares se encontraran en una misma clase. En la Figura 2 se pueden observar qué clases ejecutan qué partes del algoritmo. De esta manera, la clase NNData.java se encarga de realizar todas las tareas relacionadas con el procesamiento de datos de entrada, lectura y escritura de archivos. La clase principal NNROC.java realiza la inicialización y entrenamiento de la red neuronal usando el archivo de datos generado con la clase NNData.java. La red neuronal guarda las predicciones para la variable de respuesta evaluada y calcula la especificidad y sensibilidad del modelo. Este procedimiento se realiza de manera iterativa para cada uno de los puntos de corte definidos, calculando así todos los puntos sobre la curva ROC. Una vez obtenidos estos puntos, se estima el área bajo la curva. La clase PlotROC elabora la representación gráfica de la curva ROC y reporta tanto el área bajo la curva como el Error Estándar de la misma.

Cada una de las redes neuronales se corrió con la configuración que genera el menor RMSE. Esta configuración consta de varios parámetros que definen la red como tal, y determinan completamente su comportamiento, los cuales son: número de capas ocultas que se encuentran entre los nodos de entrada y el de salida, para este caso se determinó que se debe usar una sola capa oculta; número de neuronas en la capa oculta, tasa de aprendizaje de la red y momento. En la Tabla 2 se muestran los parámetros usados en cada uno de los modelos.

Variable de Respuesta	Neuronas en la Capa Oculta	Tasa de Aprendizaje	Momento
Discriminación Bisílabos	6	0.05847126344	0.69704828719
Discriminación Frases	9	0.22429221339	0.22159215273
Promedio Tonal	6	0.10061315455	0.77617111534

Tabla 2. Configuración de las redes neuronales.

Capítulo IV

Resultados

Este capítulo se divide en dos partes, en la primera, se muestran los resultados de las tres redes neuronales que predicen el comportamiento en las tres variables de respuesta. En la segunda parte, se realiza la comparación de medias de las áreas bajo la curva ROC de la regresión logística y de la red neuronal, donde el objetivo es determinar si existe una diferencia significativa entre estos dos modelos de predicción.

La red neuronal se corrió varias veces para cada variable de respuesta, buscando el número de puntos de corte óptimo, para el cual el área bajo la curva empezaba a converger y se minimizaba el tiempo de corrida. Se encontró que el número de puntos de corte que alcanzaba este objetivo era 20. En cada corrida se graficó la curva ROC y se estimó tanto el área bajo la curva, como su Error Estándar.

4.1. Análisis ROC para las Redes Neuronales

Discriminación de Bisílabos

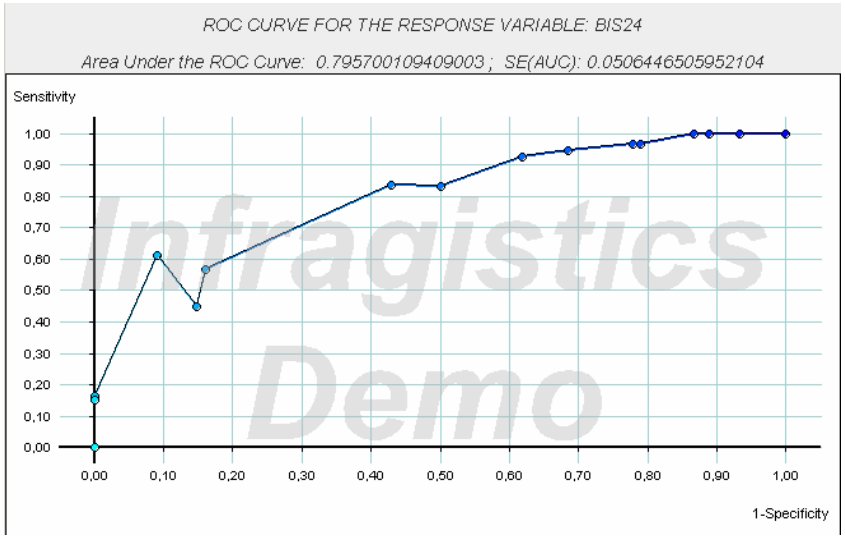


Figura 3. Curva ROC de la Red Neuronal -Discriminación de Bisílabos-

En la Figura 3 se puede observar que para esta variable de respuesta, se encontró un área bajo la curva ROC de 0.795 y un error estándar de 0.051.

Discriminación de Frases

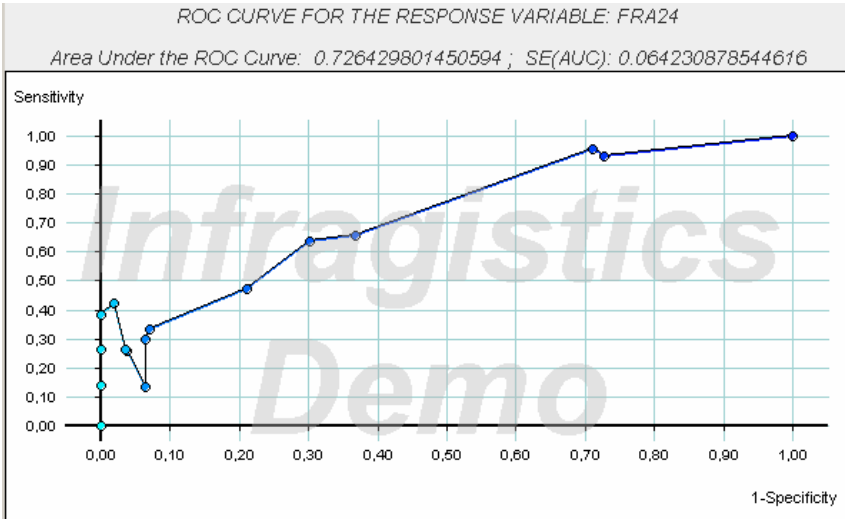


Figura 4. Curva ROC de la Red Neuronal -Discriminación de Frases-

En la Figura 4 se observa que el área bajo la curva ROC de la red neuronal es de 0.726, la cual tiene un error estándar es de 0.064.

Promedio Tonal

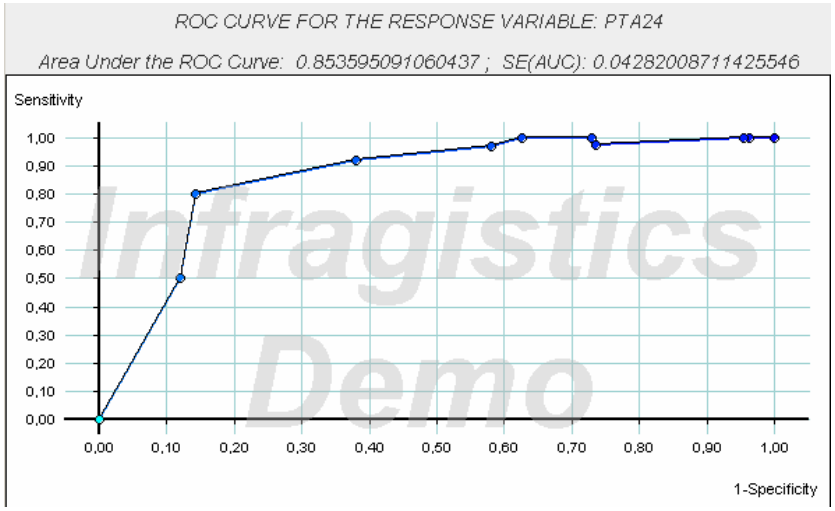


Figura 5. Curva ROC de la Red Neuronal -Promedio Tonal-

En la figura 5 se muestra que la red neuronal para la variable de respuesta promedio tonal, presenta un área bajo la curva ROC de 0.853 y un error estándar de 0.043

4.2. Red Neuronal Vs. Regresión Logística

En la Tabla 3 se presenta la comparación de las áreas bajo la curva de ambos métodos, para cada una de las variables de respuesta.

Variable de Respuesta	AUC -Regresión Logística-	SE (AUC)	AUC -Red Neuronal-	SE(AUC)
Discriminación Bisílabos	0.671	0.063	0.795	0.051
Discriminación Frases	0.532	0.070	0.726	0.064
Promedio Tonal	0.578	0.067	0.853	0.043

Tabla 3. Resumen de Resultados.

Los resultados reales de los pacientes para cada una de las variables de respuesta, usados como datos de entrada (datos contra los cuales se comparan las predicciones) son idénticos para ambos modelos predictivos. Por lo tanto, la correlación que se incluye a los modelos al usar el mismo conjunto de pacientes y los mismos datos, es igual a 1. De esta manera, se obtienen las estimaciones del error estándar para la diferencia de medias presentadas en la Tabla 4.

Variable de Respuesta	$SE(A_1 - A_2)$	Estadístico Z	P-Value
Discriminación Bisílabos	0.012	9.984	< 0.0001
Discriminación Frases	0.006	30.784	< 0.0001
Promedio Tonal	0.025	11.184	< 0.0001

Tabla 4. Error estándar para la diferencia de áreas y estadístico Z.

De acuerdo a la Tabla 4, la diferencia entre las áreas es significativa para todas las variables de respuesta, lo cual quiere decir, que la red neuronal genera siempre predicciones más precisas en comparación con las generadas por la regresión logística. En las figuras 8, 9 y 10 se muestra la superposición de las curvas ROC derivadas de las redes neuronales y de la regresión logística.

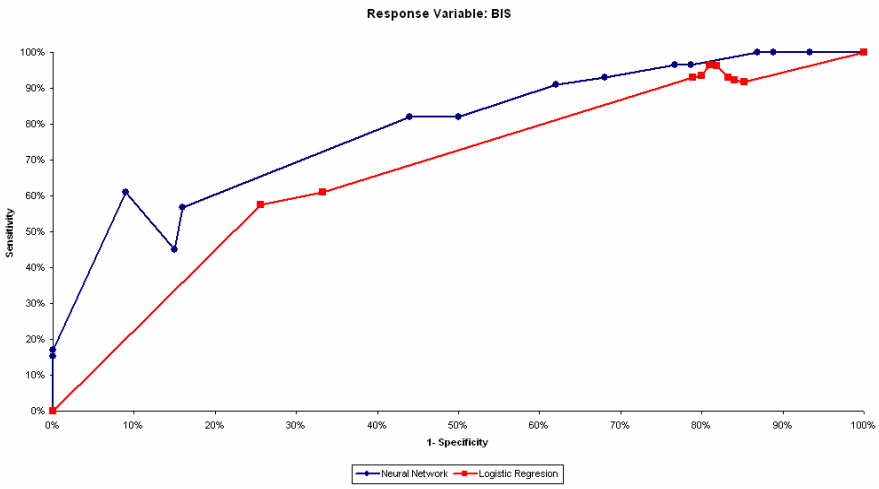


Figura 6. Comparación de las Curvas ROC -Discriminación de Bisílabos-

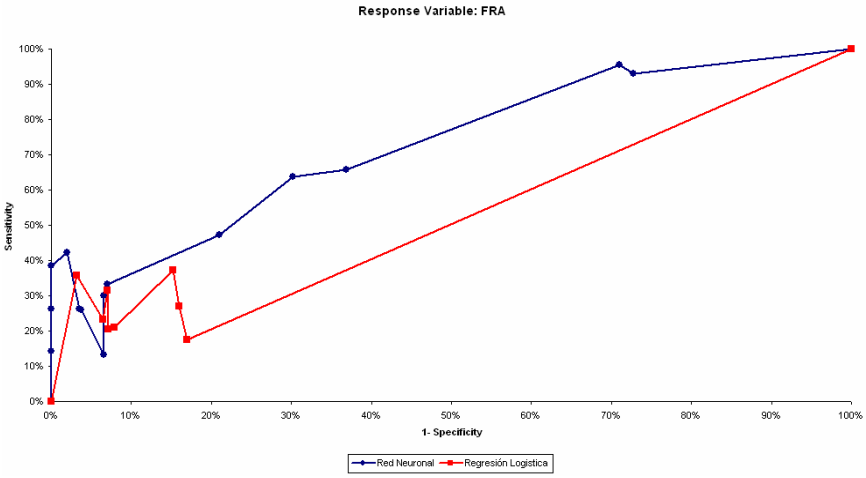


Figura 7. Comparación de las Curvas ROC -Discriminación de Frases-

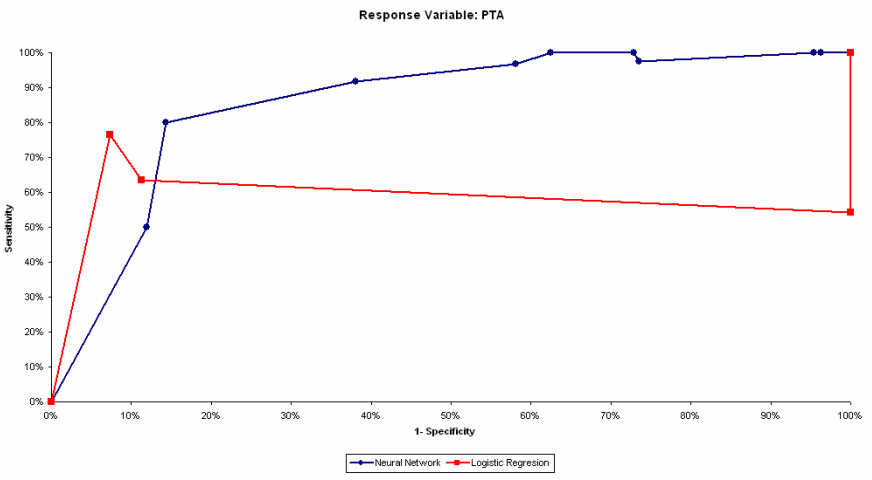


Figura 8. Comparación de las Curvas ROC –Promedio Tonal-

Capítulo V

Conclusiones e Investigación Futura

La red neuronal es más acertada que la regresión logística multivariada, para todas las variables de respuesta. Se observa que el área bajo la curva ROC, en ninguno de los casos está por encima de 0.85, lo cual indica que la precisión predictiva de la red, está de cierta manera limitada por la presencia de una cantidad significativa de observaciones ambiguas dentro de los datos de entrada. Esto quiere decir que se encuentran pacientes que tienen las mismas condiciones iniciales (valores iguales en sus variables de entrada), pero presentan resultados opuestos en las variables de respuesta. La existencia de estas observaciones ambiguas puede estar indicando, que aún no se han identificado una o varias variables de entrada significativas, que explican la diferencia de comportamiento en estos casos.

Se propone para investigación futura, usar la maximización del área bajo la curva ROC, como criterio para optimizar la configuración de las redes neuronales. El actual enfoque de minimización del RECM, puede estar limitando la capacidad predictiva del modelo en términos de su especificidad y sensibilidad [1]. Sin embargo, al implementar este criterio de optimización, se puede generar un aumento significativo en el tiempo que le tome al modelo encontrar la configuración óptima.

Adicionalmente, se puede estudiar la posibilidad de extender el uso del análisis ROC, generando una curva ROC del modelo no solo con los datos de los pacientes a los 24 meses del implante, sino también para cada paciente. En este caso, la generación de la curva ROC no se haría variando el punto de corte, sino el tiempo transcurrido después del implante, lo cual implicaría a su vez, la conformación de bases de datos a través del tiempo. Esta base de datos almacenaría la evolución de cada uno de los pacientes en las diferentes variables de entrada y de respuesta.

Referencias

[1] Rodríguez D'Alleman, Jorge H. Modelo de computación blanda para predecir la percepción del habla en niños con implante coclear. 2005

[2] Honghu L, Tongtong W. Estimating the area under a receiver operating characteristic (ROC) curve for repeated measures design.

[3] Hanley J.A, McNeil B.J The meaning and use of the area under the Receiver Operating Characteristic.1982. Radiology, 143: 30-31

[4] Hanley J.A, McNeil B.J. A method of comparing the areas under Receiver Operating Characteristic curves derived form the same cases. 1983. Radiology, 148: 840-841

[5] Holzner, Steve. Eclipse. Sebastopol, CA : O'Reilly, 2004.

[6] Eriksson, Hans-Erik. UML Toolkit. Indianapolis, Ind: Wiley Pub, 2004.

[7] JooneWorld. Java Oriented Network Engine (Joone). [Java Library]. Disponible en: http://www.jooneworld.com/. Ultimo acceso en: Abril 10, 2006.

[8] Sun Microsistems Inc. Java Technology. [Computer Program]. Disponible en http://java.sun.com/. Ultimo acceso en: Mayo 15, 2006.

[9] Ostermiller Java Utilities. Comma Separated Values (CSV). [Java Library]. Disponible en http://ostermiller.org/utils/CSV.html. Ultimo acceso en: Marzo 20, 2006.

[10] Infragistics Inc. JSuite. JFCChart [Java Library]. Disponible en: http://www.infragistics.com/ Ultimo acceso en: Abril 26, 2006.