



Universidad Nacional de La Pampa



FACULTAD DE
INGENIERÍA

Ingeniería en Sistemas

Trabajo de Tesis

**Utilización de Redes Neuronales
Artificiales y Algoritmos Genéticos para
diagnosticar enfermedades**

TESISTAS

**A.P. Delsol, Kevin
A.P. Isaia, Miguel Adrián**

DIRECTORA

Dra. Mg. Gabriela Minetti

ÁREA TEMÁTICA

Sistemas Inteligentes

General Pico, La Pampa, Argentina - 2014

Dedicatorias y Agradecimientos

Kevin Delsol

Durante el tiempo que duraron mis estudios, hubo mucha gente a mí alrededor, personas que me ayudaron y apoyaron para que hoy pueda cumplir el objetivo que me trajo a esta ciudad. Quiero agradecer a todos aquellos que, de una u otra forma, estuvieron presentes cuando los necesité. En primer lugar, agradezco a mi familia, pilar fundamental en mi desarrollo como persona, pero en especial a mi madre Graciela, por criarme, darme siempre lo que necesité, y ser quien hizo todo a su alcance para permitirme estudiar, es a ella a quien le debo mi carrera. A mi hermano Nahuel, por una vida de recuerdos, por oficiar de cocinero y personal de limpieza para que yo pudiera estudiar mi última materia, y muchas otras cosas que quedan entre hermanos. A mi padre José, con quien quizás no tenga una relación típica de padre e hijo por cuestiones de la vida, pero se hizo presente siempre que así lo necesité. A mi grupo de amigos, de mi infancia y de mi vida universitaria. A mi amigo de toda la vida Víctor Gómez, quien siempre estuvo allí, sin importar lo que hiciera falta. A mis compañeros de estudio, Nahuel Villegas, José Zanin, Darío Gette, Leonardo Appiolaza, Ezequiel Morán, Hernán Garesio, León Laborde, Darío Mattioli, Nicolás Fornione, Nahuel Ferrán, Martín Santiago y Miguel Isaia. Compañeros de estudios, madrugones, trasnoches, asados, salidas, alegrías, tristezas, pero sobre todo compañeros de la vida y mis amigos. Y gracias nuevamente a mi compañero de tesis, Miguel, ha sido una experiencia increíble trabajar en este proyecto a tu lado, llena de momentos alegres y trabajo en equipo. Por último, gracias a todos aquellos que, aunque no los haya mencionado, estuvieron ahí cuando necesite una mano y son hoy también partícipes de la conclusión de mi carrera. Muchas gracias a todos, Kevin.

Miguel Adrián Isaia

Al presente proyecto y mi carrera universitaria, se los dedico a Dios y principalmente a mi familia, que estuvieron allí en todo momento. A mi padre, a quien no dejaba escuchar la televisión debido a que yo tenía que estudiar, quien todas las mañanas funcionaba como despertador y me esperaba con la jarra de café caliente para desayunar. A mi madre que en estos 5 años de estudios me pudo mantener con lo poco que teníamos, preocupándose por mi salud, mis estudios, mi educación y alimentación entre otras cosas. A mi hermana Romina, ya que por ella escogí esta carrera y pude superar el primer año gracias a sus experiencias y apuntes. A mi hermano Matías, por compartir tantos momentos, andanzas y secretos que solo con un hermano se pueden tener, y a mi hermana Belén por prestarme su moto y darme varios consejos en esta vida universitaria. Agradecerles a mis amigos de la secundaria, que a pesar de no vernos por casi todo un año, nos encontráramos en los cumpleaños y nos escribíamos por celular, por su preocupación y todo el apoyo brindado de siempre. Una dedicatoria especial a mis dos grandes amigos Fede y Denis, de los cuales hay miles de anécdotas y travesuras para contar en un futuro a nuestros hijos. A los amigos que forme en clase, a Leonardo alias el pela, Darío alias Ñoqui, Nico alias Margarita, Hernán alias el gringo, Nahuel Ferrán y a mi compañero de tesis Kevin Delsol, solo les puedo dar las gracias por esas infinitas rondas de mates, noches de estudios, horas previas al examen de repaso, riquísimos asados, entre otras cosas. A ellos y a todos los demás que se fueron sumando. Y quiero agradecer nuevamente a Kevin, por elegirme como compañero en este último paso y por tenerme una infinita paciencia, ya que por culpa mía nos atrasamos mucho tiempo debido en parte a mi trabajo, y en otra parte a mis bajos deseos de terminar, siempre incentivándome a finalizar. Por último quisiera agradecerle a mi novia Tatiana, que ha apartado tiempo suyo para dejarme terminar este proyecto y se preocupó por mí en todo momento, y a su familia que en estos últimos tiempos me brindaron una contención enorme. Gracias a todos, Miguel.

Agradecimientos conjuntos

Agradecemos por supuesto a todo el personal de la Facultad de Ingeniería, así como también de la Universidad Nacional de La Pampa, personas que día a día trabajan para que todo aquel con deseos de estudiar pueda hacerlo. Agradecemos a la UNLPam como entidad, por facilitar, en el caso de Miguel, becas de ayuda económica, las cuales permiten que muchos alumnos no abandonen sus estudios. Gracias a todos los profesores y ayudantes de cátedra que ayudaron a formarnos como profesionales, y en especial a nuestra directora de tesis Gabriela Minetti, por asesorarnos y servir de guía en esta última etapa, sugiriendo un tema de estudio y lidiando con nuestra inexperiencia en este tipo de trabajo.

Resumen

Cuando se analiza la evolución del hardware informático, se observa un rápido crecimiento en su capacidad de cómputo y desempeño temporal. Para que sea posible un mayor aprovechamiento de dichas capacidades, es necesario el desarrollo de algoritmos cada vez más eficientes. En un principio los algoritmos no eran más que un conjunto de órdenes simples destinadas a solucionar problemas matemáticos simples, pero al crecer en capacidad las máquinas, se amplió también la gama de problemas que eran capaces de resolver.

Actualmente las características del hardware no son la principal limitación, sino el tiempo de resolución del problema. Si bien muchos problemas de la vida real pueden ser modelados y resueltos en tiempos aceptables utilizando algoritmos deterministas, hay problemas que resultan virtualmente imposibles de resolver en tiempos aceptables utilizando dichos algoritmos. Por esta razón es necesario el uso de algoritmos estocásticos. Un algoritmo determinista es aquel en el cual, conociendo las entradas del mismo, se pueden predecir sus salidas; la salida de un algoritmo estocástico, a diferencia, depende además de factores aleatorios propios del problema que modela y resuelve [17].

Considerando las restricciones temporales, una amplia gama de algoritmos estocásticos han sido desarrollados para modelar y resolver una aún más amplia gama de problemas, casos particulares son el modelado de sistemas biomédicos.

Por medio del presente trabajo de tesis, se pretende dar solución a dos problemáticas: clasificación, en benigno o maligno, de muestras de tumores de pacientes sospechados de sufrir de cáncer de mama; y clasificación, en negativo o positivo, de muestras de pacientes sospechados de padecer diabetes de tipo II. Nuestra propuesta consiste en desarrollar algoritmos basados en redes neuronales artificiales, en algoritmos genéticos y en algoritmos que hibridan estas opciones, para realizar diagnósticos confiables (clasificación) con respecto a estas enfermedades.

Tabla de Contenido

Resumen.....	IV
1. Introducción.....	9
1.1. Motivación.....	9
1.2. Organización del Trabajo de Tesis.....	11
2. Redes Neuronales Artificiales.....	13
2.1. Antecedentes Históricos.....	13
2.2. Ventajas del uso de RNAs.....	14
2.3. Cuestiones Preliminares al Trabajo con RNAs.....	15
2.4. Aplicaciones.....	16
2.5. Principios Funcionales de las Redes Neuronales Artificiales.....	17
2.5.1. Función de Entrada (Input Fuction)	20
2.5.2. Función de Activación (Activation Function)	21
2.5.3. Función de Salida (Output Function)	22
2.5.4. Niveles o Capas de una RNA.....	23
2.5.5. Mecanismos de Aprendizaje.....	24
2.5.5.1. Aprendizaje Supervisado.....	25
2.5.5.2. Aprendizaje no supervisado.....	26
2.5.6. Número de Capas Ocultas.....	27
2.5.7. Codificación de los Datos de Entrada.....	27
2.5.7.1. Codificación de los Atributos Numéricos.....	27
2.5.7.2. Codificación de los Atributos Simbólicos.....	28
2.6 Capacidad de Generalización.....	29
3. RNA Backpropagation.....	32
3.1. Estructura y Aprendizaje de una RNA Backpropagation.....	32
4. Metaheurísticas.....	38

4.1. Historia.....	38
4.2. Introducción a las Metaheurísticas.....	39
4.3. Razones para el Uso de Metaheurísticas.....	41
4.4. Conceptos Comunes a toda Metaheurística.....	42
4.4.1. Representación de las Soluciones.....	42
4.4.2. Función Objetivo.....	43
5. Algoritmos Genéticos.....	44
5.1. Algoritmos Evolutivos.....	44
5.2. Principios Funcionales de los Algoritmos Genéticos.....	44
5.2.1. Representación de las Soluciones.....	46
5.2.2. Inicialización de la Población.....	47
5.2.3. Función de Evaluación (o Función de Fitness)	47
5.2.4. Selección.....	48
5.2.4.1. Selección por Ruleta.....	48
5.2.4.2. Selección por Ranking.....	49
5.2.4.3. Selección por Torneo.....	49
5.2.5. Reproducción y Operadores de Variación.....	50
5.2.5.1. Mutación.....	50
5.2.5.2. Recombinación o Crossover.....	51
5.2.6. Reemplazo.....	53
5.2.7. Condición de Parada.....	53
5.2.8. Razones para utilizar AGs.....	54
5.2.9. Aplicaciones.....	58
6. Metaheurísticas Híbridas.....	60
6.1. Clasificación de las metaheurísticas híbridas.....	60
6.1.1. Diseño.....	61
6.1.2. Implementación.....	64

7. Enfermedades a Diagnosticar.....	66
7.1. Descripción de los conjuntos de datos utilizados.....	66
7.1.1. Conjunto de datos correspondiente al diagnóstico de cáncer de mama.....	67
7.1.2. Conjunto de datos correspondiente al diagnóstico de diabetes de tipo II.....	68
8. Diseño de los algoritmos propuestos para el diagnóstico de enfermedades.....	70
8.1. Diseño de los algoritmos basados en RNAs.....	70
8.2. Diseño de los algoritmos basados en AGs.....	74
8.3. Diseño del algoritmo híbrido propuesto.....	76
9. Experimentación y Análisis.....	78
9.1. Caso de estudio: Cáncer de mama.....	79
9.2. Caso de estudio: Diabetes.....	82
9.3. Comparación de resultados con otras investigaciones.....	84
10. Conclusiones y Trabajo Futuro.....	86
10.1. Conclusiones.....	86
10.2. Trabajos Futuros.....	87
Bibliografía.....	89
Anexo I: Interfaz Gráfica.....	93
Anexo II: Configuración paramétrica de los algoritmos propuestos.....	99

Índice de Tablas

Tabla 5.1. Correspondencia de vocablos del lenguaje natural en lenguaje de AGs.....	45
Tabla 7.1. Atributos pertenecientes al conjunto de datos de diagnóstico de cáncer de mama...	67
Tabla 7.2. Atributos pertenecientes al conjunto de datos de diagnóstico de diabetes.....	69
Tabla 8.1 Síntesis de los parámetros en cada RNA.....	73
Tabla 9.1. Acrónimos usados en los casos de estudios.....	79
Tabla 9.2. Configuración de parámetros para cada algoritmo (Cáncer de mama).....	80
Tabla 9.3. Resultados de las mejores combinaciones de parámetros (Cáncer de mama).....	80
Tabla 9.4. Configuración de parámetros para cada algoritmo (Diabetes).....	82
Tabla 9.5. Resultados de las mejores combinaciones de parámetros (Cáncer de mama).....	82
Tabla 9.6. Porcentaje de Error de Clasificación para los algoritmos.....	85
Tabla AII.1. Acrónimos usados en las tablas de experimentos.....	99
Tabla AII.2.1. Experimentación con funciones de costo y activación (RNA1-Cáncer).....	100
Tabla AII.2.2. Experimentación con funciones de costo y activación (RNA1-Diabetes).....	101
Tabla AII.3.1. Experimentación con los parámetros β y η (RNA1-Cáncer).....	101
Tabla AII.3.2. Experimentación con los parámetros β y η (RNA1-Diabetes).....	102
Tabla AII.4.1. Experimentación con épocas de entrenamiento y cantidad de neuronas en la capa oculta (RNA1-Cáncer).....	102
Tabla AII.4.2. Experimentación con épocas de entrenamiento y cantidad de neuronas en la capa oculta (RNA1-Diabetes).....	103
Tabla AII.5.1. Experimentación con épocas consideradas para ajustar el η y valor del parámetro α (RNA2-Cáncer).....	104
Tabla AII.5.1. Experimentación con épocas consideradas para ajustar el η y valor del parámetro α (RNA2-Diabetes).....	104
Tabla AII.6.1. Experimentación con los parámetros a y b (RNA2-Cáncer).....	105
Tabla AII.6.2. Experimentación con los parámetros a y b (RNA2-Diabetes).....	105
Tabla AII.7.1. Experimentación con cantidad de generaciones y tamaño de población (AG1-Cáncer).....	107
Tabla AII.7.2. Experimentación con cantidad de generaciones y tamaño de población (AG1-Diabetes).....	107

Tabla AII.8.1. Experimentación con la probabilidad de mutación (AG1-Cáncer).....	108
Tabla AII.8.2. Experimentación con la probabilidad de mutación (AG1-Diabetes).....	108
Tabla AII.9. Experimentación con AG2.....	109
Tabla AII.10.1. Experimentación con el híbrido AGRNA (Cáncer).....	110
Tabla AII.10.2. Experimentación con el híbrido AGRNA (Diabetes).....	110

Índice de Ilustraciones

Figura 2.1. Comparación entre una neurona biológica (izquierda) y una artificial (derecha).	19
Figura 2.2. Neurona artificial con dos entradas y una salida	19
Figura 2.3. Esquema básico de una RNA.....	20
Figura 2.4 Ejemplo de una Red Neuronal multicapa.....	24
Figura 2.5. Influencia de la salida de la Neurona N_j en la entrada de la Neurona N_i	26
Figura 2.6. Transformación de la edad al intervalo $[0.1, 0.9]$	28
Figura 2.7. Transformación de la edad en tres atributos simbólicos.....	29
Figura 3.1. Uso del bias en la capa X de una RNA	34
Figura 5.1. Ejemplo Selección por Ruleta.....	49
Figura 5.2. Ejemplo de mutación de un gen dentro de un cromosoma	51
Figura 5.3. Crossover de un punto.....	52
Figura 5.4. Crossover de dos puntos	52
Figura 6.1. Clasificación de la hibridación según el diseño.....	61
Figura 6.2. Clasificación de la hibridación según la implementación.....	64
Figura 8.1. Esquema general de las RNAs desarrolladas.....	71
Figura 8.2. Adición del bias a la estructura de la red.....	72
Figura 8.3. Matrices de pesos sinápticos para las RNAs.....	72
Figura 8.4. Composición de cada individuo de la población de un AG.....	75
Figura 8.5. Mutación de un individuo por medio de la RNA.....	77
Figura 9.1. Boxplots y comparación múltiple de los valores de EME/EMI obtenidos por cada algoritmo.....	81
Figura 9.2. Boxplots y comparación múltiple de los valores de EMT obtenidos por cada algoritmo.....	81
Figura 9.3: Boxplots y comparación múltiple de los valores de EME/EMI obtenidos por cada algoritmo.....	83
Figura 9.4: Boxplots y comparación múltiple de los valores de EMT obtenidos por cada algoritmo.....	84
Figura AI.1. Panel correspondiente a las RNAs.....	93
Figura AI.2. Panel correspondiente a los AGs.....	95
Figura AI.3. Panel correspondiente a AGRNA.....	95
Figura AI.4. Panel para la comparación de datos.....	96

Figura AI.5. Sistema en tiempo de ejecución. Ejecutando una RNA.....	97
Figura AI.6. Sistema en tiempo de ejecución. Ejecutando un AG.....	97
Figura AI.7. Sistema en tiempo de ejecución. Ejecutando el híbrido AGRNA.....	98
Figura AII.1. Progresión del error de entrenamiento (RNA1-Cáncer)	106
Figura AII.2. Progresión del error de entrenamiento (RNA1-Diabetes)	106
Figura AII.3. Evolución del error del mejor individuo de la población (AG2-Cáncer)	109
Figura AII.4. Evolución del error del mejor individuo de la población (AG21-Diabetes)	109
Figura AII.5. Evolución del error del mejor individuo de la población (AGRNA-Cáncer) ...	111
Figura AII.6. Evolución del error del mejor individuo de la población (AGRNA-Diabetes).	111

Índice de Ecuaciones

Ecuación 2.1. Función de Entrada de una RNA.....	20
Ecuación 2.2. Función de Entrada usando Sumatoria.....	21
Ecuación 2.3. Función de Entrada usando Producto.....	21
Ecuación 2.4. Función de Entrada usando Máximo.....	21
Ecuación 2.5. Función lineal.....	22
Ecuación 2.6. Función Sigmoidea.....	22
Ecuación 2.7. Función Tangente Hiperbólica.....	22
Ecuación 2.8. Función de Salida Binaria.....	23
Ecuación 2.9. Calculo del peso sináptico.....	25
Ecuación 2.10 regla de aprendizaje de un perceptrón.....	26
Ecuación 2.11. Codificación de atributos numéricos.....	28
Ecuación 3.1. Cálculo de entrada de una neurona oculta.....	33
Ecuación 3.2. Cálculo de las salidas de las neuronas ocultas.....	34
Ecuación 3.3. Cálculo de las salidas de las neuronas de salidas.....	34
Ecuación 3.4. Cálculo del error de las neuronas.....	34
Ecuación 3.5. Derivada de una función lineal.....	35
Ecuación 3.6. Derivada de una función Sigmoidea.....	35
Ecuación 3.7. Error con medida entrópica.....	35
Ecuación 3.8. Corrección de pesos de las neuronas de la capa de salida	36
Ecuación 3.9. Corrección de pesos de las neuronas de la capa oculta.....	36
Ecuación 3.10. Corrección de pesos con momentum.....	36
Ecuación 3.11. Eta adaptativo.....	37