Desarrollo de una plataforma astroinformática para la administración y análisis inteligente de datos a gran escala

Clasificación de Galaxias mediante Redes Neuronales Artificiales

Ricardo Contreras, Gabriel Salazar, Neil Nagar,
Mauricio Solar, Marcelo Mendoza
Jorge Ibsen, Lars Nyman, Eduardo Vera, Diego Mardones, Guillermo Cabrera,
Paola Arellano, Karim Pichara, Nelson Padilla, Victor Parada.

Concepción, 5 de mayo de 2014

Resumen

El presente documento muestra el estado de avance del proyecto de minería de datos basado en la clasificación de galaxias; tema en el cual está inserta la propuesta efectuada al proyecto Fondef. Se muestra la solución propuesta para el problema que se aborda, además de la ejecución de las pruebas efectuadas, junto con un análisis de los resultados obtenidos durante éstas. Los resultados obtenidos nos muestran la capacidad de la red neuronal de reproducir la clasificación humana, llegando a obtener en el mejor resultado de las pruebas un 98 % de éxito en la tarea de clasificación.

Palabras Claves: Galaxias, Clasificación de Galaxias, Galaxy Zoo, SDSS, ChiVO, Minería de Datos, Machine Learning, Redes Neuronales Artificales.

1. Resumen Ejecutivo

Desde que Edwin Hubble en el año 1926 ideó su famosa secuencia, ha surgido la necesidad de clasificar las galaxias según ciertas características particulares. Con el tiempo la cantidad de galaxias descubiertas ha aumentado enormemente, elevando así la complejidad en términos de tiempo a ocupar para clasificarlas. Debido a esto, surge el problema de clasificar las galaxias en un tiempo razonable y con cierto grado de precisión. Las máquinas de aprendizaje cobran aquí gran valor, debido a sus características propias, ya sea de emulación de las neuronas cerebrales o la imitación del comportamiento de seres vivos.

La organización del informe será como sigue, primero se definirá el problema a abordar junto con los objetivos. Posteriormente se procederá a mostrar la solución propuesta, junto a las pruebas y sus resultados. Finalmente se mostrará el análisis de los resultados obtenidos durante las pruebas de las redes neuronales, además de las conclusiones y trabajos futuros sobre el tema desarrollado.

Índice

2. Definición del Problema

2.1. Contexto del Problema

Desde que Edwin Hubble en el año 1926 ideó su famosa secuencia [6], ha surgido la necesidad de clasificar las galaxias según ciertas características particulares. Con el tiempo la cantidad de galaxias descubiertas ha aumentado enormemente, elevando así la complejidad en términos de tiempo a ocupar para clasificarlas. Debido a esto, surge el problema de clasificar las galaxias en un tiempo razonable y con cierto grado de precisión. Las máquinas de aprendizaje cobran aquí gran valor, debido a sus características propias, ya sea de emulación de las neuronas cerebrales o la imitación del comportamiento de seres vivos.

La memoria de título intenta ser la base para una investigación futura de clasificación de fusiones de galaxias mediante máquinas de aprendizaje automático, investigación que es necesaria para entender el comportamiento de las galaxias así como también, por ejemplo, la creación de agujeros negros.

La memoria de título forma parte del proyecto de Observatorio virtual para Chile (CHIVO), existente desde el año 2013, y en el cual trabajan 5 Universidades (Universidad Técnica Federico Santa María, Universidad de Chile, Universidad de Santiago de Chile, Pontificia Universidad Católica de Chile y la Universidad de Concepción). Este Observatorio virtual intenta ser una plataforma virtual de acceso para los astrónomos, a los datos existentes en los observatorios ubicados dentro de territorio chileno. Se está trabajando en distintos tópicos como bases de datos, desarrollo de la plataforma web del proyecto, semántica de datos y minería de datos. Esta memoria está enmarcada dentro del contexto de minería de datos de CHIVO, parte fundamental para explorar e investigar los datos que se generen de una eventual alianza con los observatorios físicos nacionales y también con el Observatorio Virtual Internacional, y sus alianzas.

En el contexto real, en el futuro, CHIVO procesará y analizará los datos del observatorio ALMA (Atacama Large Millimeter Array), tendrá un repositorio de imágenes tan grande que se necesitará optimizar los tiempos en las distintas tareas que realizarán, donde la clasificación de galaxias, punto central de esta memoria de título, ocupa un lugar importante a tener en cuenta.

El clasificar galaxias es también necesario en otros problemas; además de clasificarlas según el esquema de Hubble, permite detectar fusiones de galaxias y otro tipo de morfologías no apreciables en el esquema anterior. Esta potencialidad es sumamente relevante para futuras investigaciones astronómicas, ya que ayudaría a agilizar algunas tareas donde es imposible realizarlas manualmente, por los tiempos necesarios para su procesamiento. En esta Memoria de Título se necesitó manejar un volumen de datos muy grande, por lo que el manejo de esta información es esencial para las pretensiones de ésta.

El problema entonces, se centra en clasificar galaxias de manera rápida, utilizando para ello una máquina de aprendizaje automático como es el caso de las redes neuronales artificiales.

2.2. Objetivos

2.2.1. Objetivo General

- Clasificar imágenes del catálogo de galaxias de Galaxy Zoo, versión SDSS Release 7.
- Establecer las bases que permita en un futuro clasificar fusiones de galaxias ocupando, junto con las imágenes, el Redshift (Corrimiento al rojo) u otro tipo de características no ocupadas en investigaciones anteriores, como filtro de precisión.

2.2.2. Objetivos Específicos

- Clasificar Galaxias según su morfología ocupando redes neuronales.
- Investigar sobre los distintos tipos de redes neuronales artificiales con los cuales se puede hacer la clasificación.
- Establecer una buena configuración de la red neuronal que mejore los resultados obtenidos por trabajos anteriores.
- Evaluar el prototipo diseñado y creado.

3. Solución propuesta

La resolución del problema propuesto mediante redes neuronales artificiales requiere la definición de las componentes de la solución propuesta. A continuación se muestra cada componente del modelo a desarrollar.

3.1. Conjunto de Datos

De acuerdo a la investigación realizada se tomó la decisión de manejar la muestra de datos de Galaxy Zoo, ocupada también por Banerji et al. [2]. Esta elección es motivada por la gran cantidad de datos clasificados por seres humanos disponibles en GZ1, junto con la disposición de probar variantes al problema desarrollado por Banerji et al.[2]. La clasificación de GZ1 se encuentra disponible libremente en su sitio web[1], en la cual disponen de siete tablas de datos, cada una con características distintas. Para efectos de este trabajo se ocupó la tabla número 7, la cual tiene la clasificación completa del proyecto GZ1, junto con las probabilidades que se obtuvieron del estudio exhaustivo realizado por Bamford et al. [1] al sesgo de la clasificación para poder clasificar cada imagen en su categoría correspondiente. Esta tabla contiene la fracción de votos en cada una de las seis categorías[2], en las que los usuarios de Galaxy Zoo votaron por cada imagen presentada ante ellos. Esta fracción de votos se ocupó como la salida deseada de la red durante el entrenamiento, y se representó como un vector de la forma [espiral[3],elíptica,fusión,Objeto desconocido/estrella], el cual representa la fracción de voto en cada categoría. Por ejemplo, la galaxia con el ObjID[4] 587730774962536596 tiene una fracción de votos de 0,077 en la categoría Espiral, de 0,885 en la categoría Elíptica, de 0,019 en la categoría Fusión, y de 0,019 en la categoría Objeto desconocido/estrella, por lo que el vector de esta galaxia estaría dado por $[0{,}077, 0{,}885, 0{,}019, 0{,}019]$.

Como Galaxy Zoo 1 ocupó el Data Release 7 de SDSS para su clasificación, se necesitó trabajar con los datos de este catálogo. Los datos de SDSS DR7 están almacenados en una base de datos a la cual se puede acceder desde la interfaz web de CasJobs, o bien mediante un cliente implementado en lenguaje JAVA. La obtención de los datos se hizo a través de CasJobs (Figura 1), mediante una consulta en lenguaje SQL. Los parámetros elegidos para servir de entrada a la red fueron los ocupados en el trabajo de Banerji et al. [2], correspondientes a los parámetros basados en el color y perfiles ajustados de brillo, y a los basados en forma y textura de la galaxia. La consulta efectuada extrajo los parámetros de cada galaxia clasificada por GZ1, eliminó las galaxias que contenían datos nulos o espurios, y filtró los datos para algunas pruebas especificas que se realizaron. Los datos consultados se extrajeron de la tabla `PhotoObjAll` de CasJobs y se almacenaron en el formato CSV[5].

Por otro lado, al querer entrenar las redes, con los datos de la tabla 7 de Galaxy Zoo sin ningún tratamiento posterior, vimos que no genera un entrenamiento adecuado. Esto se pudo verificar mediante algunas pruebas preliminares, las cuales arrojaron una gran cantidad de falsos positivos en la clasificación. Para solucionar el problema, se procedió a seleccionar los objetos que contenían una fracción de votos igual

[1] http://data.galaxyzoo.org/

[2] Elípticas, Espirales con brazos en sentido horario y anti-horario, espirales barradas, fusiones y objetos desconocidos/estrellas

[3] Suma de la fracción de todos los tipos de galaxias espirales (con brazos en sentido horario, anti-horario y barradas).

[4] ObjID es el identificador de la galaxia en el DR7 de Sloan.

[5] Valores separados por coma

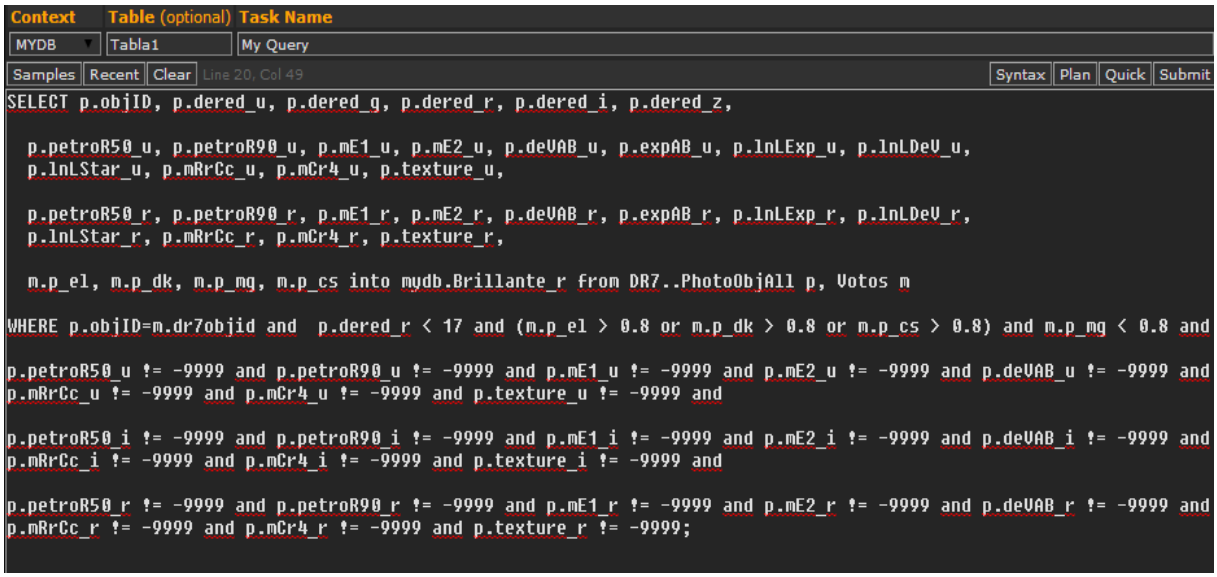

Figura 1: Ejemplo de consulta en la interfaz web de CasJobs

o mayor a 0,8 en alguna de las categorías -ya sea espiral, elíptica, fusión u objeto desconocido/estrella- y se escalarizó el vector que contiene la clasificación en forma de fracción de voto. Por ejemplo, si un objeto tiene un vector [0,89, 0,07, 0,02, 0,02], y se escalariza, el nuevo vector sería [1,0,0,0], lo que clasificaría según los votos de GZ1 al objeto como galaxia espiral.

Posteriormente se procedió a calcular la Elipticidad adaptativa (aE) y la variable de Concentración basada en los parámetros Petrosian, dado que estos datos no los entrega directamente SDSS DR7. Todas estas tareas de pre-procesamiento se llevaron a cabo con la herramienta de software matemático MATLAB.

Cabe destacar que esta muestra generada coincide con la muestra `CLEAN` propuesta por Lintott et al.[8] en su artículo de la presentación del catálogo de GZ1, y se puede comparar en parte a la muestra "Gold Sample" ocupada por Banerji et al.[2]. La cantidad de objetos de la selección efectuada fue de 287,897, donde el $34,2\,\%$ son consideradas galaxias espirales, $64,53\,\%$ galaxias elípticas y el $1,25\,\%$ son objetos desconocidos/estrellas. Las fusiones de galaxias no fueron consideradas en el estudio. Por otra parte el umbral de fracción de votos de 0,8 otorga una confiabilidad aceptable en la clasificación efectuada por los usuarios de GZ, por lo tanto es ideal para la realización de las pruebas requeridas. Por lo mostrado en Banerji et al.[2], existe otra manera de elegir el umbral de esta clasificación, pero no es el objetivo de esta memoria ahondar en ello. También se consideró probar la red neuronal con otro set de datos, cuya diferencia con el set de datos anterior, es que se consideran los objetos más brillantes. Esto quiere decir que al igual que el set de datos con objetos brillantes ocupado por Banerji et al.[2], se consideró la magnitud del filtro $r < 17$ para seleccionar los objetos. La cantidad de objetos fue de 157,918, con un $42,1\,\%$ de galaxias espirales, un $56,65\,\%$ de galaxias elípticas y un $1,25\,\%$ de objetos desconocidos/estrellas. Se realizó una sola prueba con este set con la intención de validar el entrenamiento con el set menos brillante.

En tanto, los filtros elegidos para realizar las pruebas fueron los filtros u y r. El filtro u fue elegido considerando lo mostrado en Kennicutt [7], de donde se infiere que las galaxias elípticas generalmente tienen una mayor diferencia entre la magnitud del filtro u y el filtro g, lo que ocurre también entre u

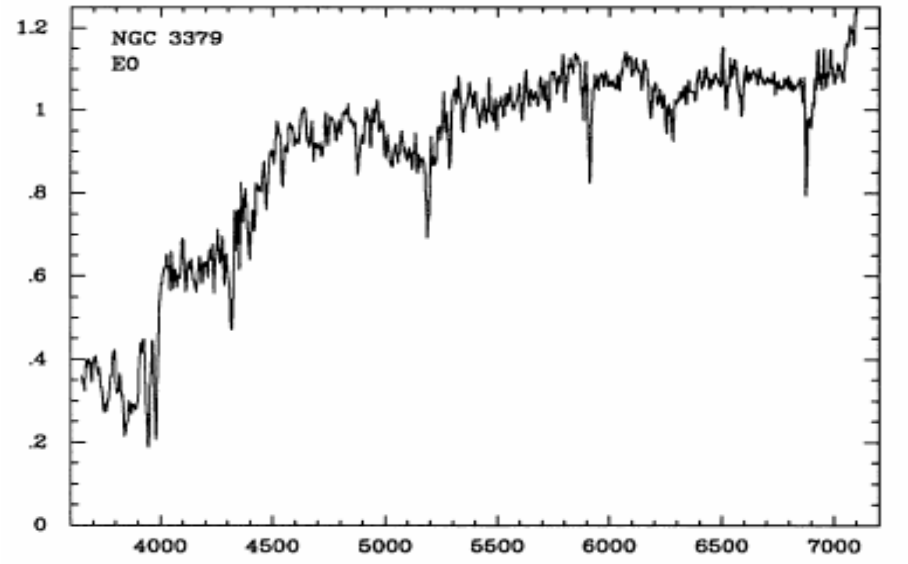

Figura 2: Espectro de la galaxia elíptica de tipo Hubble E0, NGC 3379.

Figura 3: Galaxia elíptica NGC 3379.

y los demás filtros. Como se puede ver en la figura2, la galaxia elíptica NGC[6] 3379 (Figura 3) tiene un crecimiento brusco entre los 4000$\mathring{A}$ y 4500$\mathring{A}$, que corresponde a los filtros u y g. En cambio en una galaxia espiral esto no ocurre así, se puede observar una diferencia menor entre el filtro u y los otros filtros. Esto se puede observar en el espectro (Figura 4) de la galaxia espiral NGC 4775 (Figura 5), en el cual se ve un espectro con diferencias menores entre los rangos de los filtros.

[6]El *Nuevo Catálogo General de Nebulosas y Cúmulos de Estrellas*, conocido como NGC, es un catálogo astronómico compilado en el año 1880 por John Louis Emil Dreyer en base a las observaciones realizadas años antes por William Herschel.

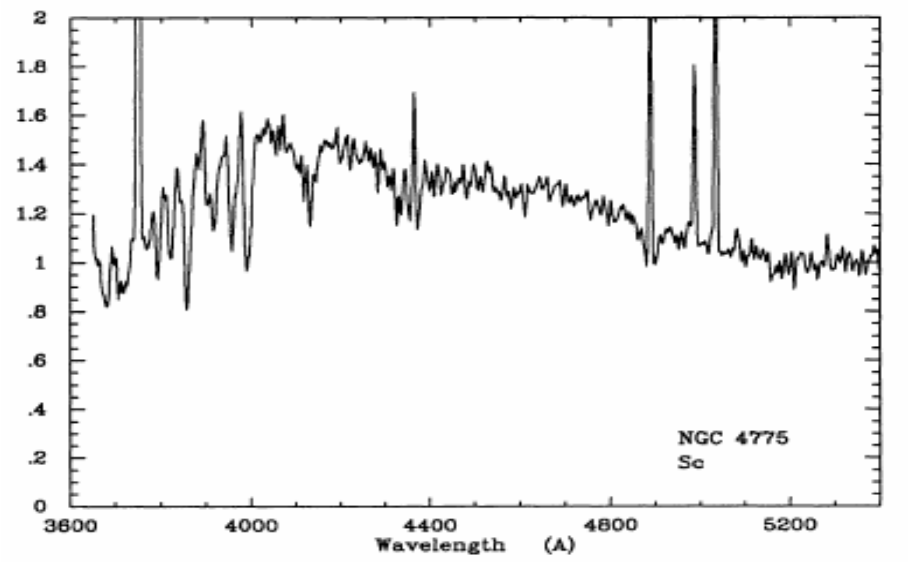

Figura 4: Espectro de la galaxia espiral de tipo Hubble Sc, NGC 4775.

Figura 5: Galaxia espiral NGC 4775.

La elección del filtro r en cambio, se basó en hacer pruebas con un filtro cercano al filtro i ocupado por Banerji et al.[2], y que estuviera más al centro en el espectro del objeto astronómico.

3.2. Entorno desarrollador

Existe una inmensa variedad de herramientas y entornos desarrolladores para implementar y probar redes neuronales artificiales, y dentro de todos estos se analizaron dos: Neurosolutions y MATLAB. El primero es un software pagado, que cuenta con una versión "Lite" (no pagada) con algunas restricciones a la hora de obtener algunos resultados del entrenamiento de las redes neuronales. Cuenta con múltiples opciones de implementación, que van desde redes con Perceptrón multicapa FeedForward, redes Recurrentes hasta los mapas auto-organizativos de Kohonen, por nombrar sólo algunas. Cuenta con una interfaz amigable, con un asesor a la hora de querer probar alguna red específica, pero que no permite cambiar de manera intuitiva algunas opciones de implementación, en el caso de que la red no esté entregando los resultados esperados. En sus últimas actualizaciones ha implementado el procesamiento paralelo a través de GPU[7], con lo cual ha hecho disminuir el tiempo de entrenamiento de las redes. Cabe destacar que esta última función sólo está disponible en la versión pagada.

MATLAB por su parte ofrece al usuario, una herramienta completa para implementar redes neuronales artificiales. Cuenta con una aplicación llamada `nprtool` la cual guía a resolver problemas tales como el reconocimiento de patrones, entre otros. También cuenta con un editor que permite al usuario modelar una red neuronal de manera sencilla, donde se puede configurar: el tipo de red neuronal, la cantidad de datos de entrenamiento, de validación y de testing, el tipo de función de activación, el algoritmo de aprendizaje, la cantidad de épocas (ciclos) de entrenamiento, el gradiente mínimo, el máximo número de errores de validación, además de las salidas de los indicadores de rendimiento de la red. Por otro lado, MATLAB cuenta con la opción de procesar las redes neuronales de forma paralela, permitiendo así un rápido entrenamiento y ahorro de tiempo importante. Todo este conjunto de variables hace de MATLAB una herramienta flexible a la hora de implementar una red neuronal. Debido a las razones expuestas se eligió MATLAB como entorno desarrollador.

[7] Unidad de procesamiento Gráfico

4. Pruebas

4.1. Configuración del entrenamiento

Tomando en cuenta el trabajo efectuado por Banerji et al. [2], se decidió emular parte de la implementación efectuada por ellos. La configuración de la red se fijó en $n : 24 : 24 : 3$, donde n es el número de parámetros de entrada para cada prueba, con dos capas ocultas de 24 neuronas cada una y tres neuronas de salida. El tipo de red neuronal elegida fue el Perceptrón multicapa, junto con el algoritmo de aprendizaje de quasi-Newton. La elección del algoritmo se debió a la limitación de memoria del equipo ocupado para las pruebas, ya que en primera instancia se había optado por el algoritmo de Levenberg-Marquard, pero como se mostró en el capitulo 1, a mayor cantidad de patrones de aprendizaje, mayor gasto de memoria efectúa. La función de activación elegida fue la función sigmoidal para las capas ocultas, y la función identidad para la capa de salida. La elección del modelo de arquitectura, que en nuestro caso fue Feedforward, influye en la elección de las funciones de activación, ya que MATLAB al elegir este tipo de arquitectura toma por defecto las funciones ya descritas. En tanto, los parámetros de detención para cada entrenamiento fueron los siguientes:

- Gradiente Mínimo: 10^{-6}
- Máximo número de comprobaciones de validación[8]: 10 comprobaciones
- MSE Objetivo: 0
- Número de épocas máxima: 1000

La red detendrá el entrenamiento cuando se cumpla uno de los criterios anteriores primero, ya sea porque se alcanzó el error mínimo, se alcanzó el número arbitrario de ciclos (épocas) elegido, se llegó al gradiente mínimo elegido, o se alcanzó el número máximo de comprobaciones de validación para prevenir el sobre-entrenamiento (Overfitting). Se normalizaron los parámetros de acuerdo a la función de MATLAB `mapstd`, la cual normaliza los datos en base a la varianza unitaria y a la media entre todos los datos de cada parámetro de entrada de la red. También se probó la normalización `mapminmax`, que transforma los datos de los parámetros al rango $[-1, 1]$, pero no hubo mayor diferencia de rendimiento con respecto al otro método durante el entrenamiento. La división de los datos de entrenamiento, de validación y de testing se hizo de acuerdo a la función `divideint`, la cual divide los datos utilizando bloques intercalados aleatorios. El porcentaje elegido para cada set de datos fue el siguiente: un 85 % del total de datos para el entrenamiento, un 10 % para la validación y un 5 % para el testing. Se consideró esta elección para los dos conjuntos de prueba descritos en el capitulo anterior.

El rendimiento se midió mediante el MSE o *Error Cuadrático Medio*. El objetivo del MSE es minimizar el error entre la salida deseada y la salida que otorga el entrenamiento, y está definido como sigue:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(e_i)^2,$$

[8]Número máximo de comprobaciones sucesivas en donde el error del rendimiento del set de validación durante el entrenamiento no decrece.

donde e_i es la diferencia entre la salida deseada y la salida de la red y N es el número total de objetos ocupados durante el entrenamiento.

Por otra parte, se decidió también ocupar otro método de medición de rendimiento para validar las redes neuronales, la matriz de confusión [9]. Esta matriz muestra la cantidad de objetos bien clasificados durante el entrenamiento, validación y testing, por cada clase. Un ejemplo de esta matriz para dos clases se muestra en la tabla 1. En el ejemplo el valor a representa la cantidad de clasificaciones de Clase 1 bien clasificadas, el valor d la cantidad de clasificaciones de Clase 2 bien clasificadas, y los valores c y b representan la cantidad de clasificaciones mal efectuadas por la red, con respecto al valor deseado o esperado.

	Deseado Clase 1	Deseado Clase 2
Predicho Clase 1	a	c
Predicho Clase 2	b	d

Cuadro 1: Matriz de Confusión - Ejemplo

También se considera el porcentaje de éxito, el cual mide qué tan precisa es la red. Éste se calcula considerando la matriz de confusión de la red. Tomando el ejemplo de la tabla 1, su porcentaje de éxito se calcula como sigue:

$$\%\, de\, éxito = \frac{a+d}{a+b+c+d} \times 100.$$

Por ultimo, tomando en cuenta la cantidad de clasificaciones mal efectuadas de la matriz de confusión, se calcula el porcentaje de error de la red. La siguiente ecuación muestra el cálculo de este porcentaje:

$$\%\, de\, error = \frac{b+c}{a+b+c+d} \times 100.$$

4.2. Plan de pruebas

Para ver el rendimiento de las redes neuronales en el problema objetivo, se procedió a diseñar un plan de pruebas, ocupando la información de los filtros fotométricos u y r. Estos filtros se eligieron ya que se quiso variar el problema resuelto en Banerji et al.[2] y ver qué tan efectivas son las redes neuronales en otros filtros fotométricos distinto al filtro i ocupado por ellos.

Se idearon ocho pruebas, tomando distintos números de parámetros y distintos filtros. Estas pruebas se muestran a continuación:

- **Prueba 1**: Parámetros basados en el color y en los perfiles ajustados, en el filtro u.
- **Prueba 2**: Parámetros basados en la forma y textura, en el filtro u.
- **Prueba 3**: Los parámetros de Prueba 1 y Prueba 2 juntos, en el filtro u.
- **Prueba 4**: Parámetros basados en el color y en los perfiles ajustados, en el filtro r.
- **Prueba 5**: Parámetros basados en la forma y textura, en el filtro r.
- **Prueba 6**: Los parámetros de Prueba 4 y Prueba 5 juntos, en el filtro r.
- **Prueba 7**: Los parámetros de Prueba 4 y Prueba 5 juntos, en el filtro r, pero considerando los objetos brillante, es decir, tomando sólo los objetos con un $r < 17$[9].
- **Prueba 8**: Los parámetros de Prueba 4 y Prueba 5, pero eliminando el parámetro *texture*, en el filtro r.

Las pruebas con el filtro u (**Pruebas 1**, **2** y **3**), se eligieron para probar el rendimiento de la red con las dos clases de parámetros propuestos. En este caso se quiere probar si el filtro u, debido a su naturaleza, clasifica bien las galaxias espirales y elípticas. Por otro lado, las pruebas con el filtro r (**Pruebas 4**, **5** y **6**) deben servir para probar si ocupando ese filtro, se obtienen tan buenos resultados como los mostrados en Banerji et al.[2] ocupando el filtro i. Estos últimos resultados no se pueden comparar directamente, ya que Banerji et al. ocupó un sesgo un poco más complejo que el ocupado aquí, ademas de ocupar una variación de la matriz de confusión al entregar los resultados.

Se agregó la prueba de eliminar el parámetro *texture* (**Prueba 8**) y la prueba con el conjunto de objetos brillantes (**Prueba 7**), para mostrar cómo afectan cada una el rendimiento de la red, y compararlas con la **Prueba 6**. La idea de eliminar el parámetro *texture* proviene de analizar el histograma de distribución de los tipos de galaxias en el parámetro en cuestión, mostrados en Banerji et al.[2]. Se puede apreciar que la distribución de los objetos no difiere mucho al considerar *texture*, y lo que se busca analizar es la necesidad de mantener este parámetro o no.

Para cada prueba se procedió a realizar 10 ejecuciones[10] de la red, eligiéndose la última de ellas. Esto debido al análisis de los valores obtenidos en cada medidor de rendimiento entre las diez ejecuciones, y se llegó a la conclusión de que no diferían demasiado entre cada ejecución como para afectar el rendimiento global de la prueba.

[9] El tamaño de este subconjunto es de 157,919 objetos, cantidad menor a la considerada en las otras pruebas.

[10] Una ejecución corresponde al entrenamiento, validación y testing de una prueba de la red neuronal.

5. Resultados

La presentación de los resultados están organizados de la siguiente forma. En primer lugar se presentan los porcentajes de éxito y de error, MSE alcanzado, el motivo de la detención del entrenamiento, el número máximo de épocas alcanzado, y el tiempo de ejecución de cada prueba. Por otra parte se presenta la matriz de confusión junto al gráfico de comparación, el cual compara la cantidad de objetos de cada clase clasificados por la red neuronal contra la cantidad de objetos clasificados por los usuarios de Galaxy Zoo 1. Debido a esto se puede apreciar el grado de acercamiento a la clasificación objetivo (deseada) por parte de cada prueba de la red neuronal. Los resultados de la ejecución de cada prueba se muestran en el documento Memoria de Título y el resumen de ellas, sin las matrices de confusión, se muestra a continuación.

5.1. Resumen

En esta sección se mostraran las tablas resumen de las pruebas efectuadas. Las tablas se presentarán separadas en tres tipos, dependiendo de las características de cada prueba efectuada para su posterior análisis. Cada tipo se describe a continuación.

- Tipo A: Toma las pruebas 1, 2 y 3, las cuales tienen en común el filtro fotométrico u, y se diferencian por los parámetros de entrada a la red.

	Filtro u		
	Prueba 1	Prueba 2	Prueba 3
% de éxito	94,3	89,8	95,9
% de error	5,7	10,2	4,1
MSE	0,029	0,053	0,0215

Cuadro 2: Resumen de Pruebas Tipo A

- Tipo B: Toma las pruebas 4, 5 y 6, que son similares a las pruebas del Tipo A, pero que ocupan la información del filtro r.

	Filtro r		
	Prueba 4	Prueba 5	Prueba 6
% de éxito	97,4	95,6	98,0
% de error	2,6	4,4	2,0
MSE	0,0133	0,0227	0,0105

Cuadro 3: Resumen de Pruebas Tipo B

- Tipo C: Son de este tipo las pruebas 7 y 8, ya que son variaciones a la prueba 6 del Tipo B.

	Filtro r	
	Prueba 7	Prueba 8
% de éxito	97,5	98,0
% de error	2,5	2,0
MSE	0,013	0,010

Cuadro 4: Resumen de Pruebas Tipo C

.

6. Análisis de Resultados

El análisis de resultados de un experimento conlleva a evaluar las similitudes y diferencias entre cada par de pruebas efectuadas. En ese contexto, se procedió a agrupar las pruebas en tres tipos, dependiendo de sus características particulares.

Uno de los primeros análisis, fue ver las diferencias entre las pruebas de Tipo A y Tipo B. Como se puede apreciar en las tablas 2 y 3, las pruebas de Tipo B resultaron mejores que las de Tipo A, significando así que las pruebas realizadas ocupando los parámetros asociados al filtro r fueron superiores a todas las pruebas que ocuparon los parámetros del filtro u. Lo que nos lleva a resumir que la red se comporta de mejor manera con el filtro fotométrico r.

Por otro lado, comparando las pruebas individuales dentro de su mismo tipo, es decir, viendo las pruebas que tienen distintos tipos de parámetros elegidos para el entrenamiento independiente del filtro ocupado, se obtiene lo siguiente. Para las pruebas de Tipo A (Filtro u), la prueba 3 fue la que mejor comportamiento tuvo, sin embargo, la prueba 2 tuvo un pobre desempeño, incluso comparándola con todas las demás pruebas. Esto dice, que por si solos, los parámetros basados en la forma y textura de la galaxia no son tan buenos parámetros en comparación con los basados en el color y perfiles ajustados del objeto. Sin embargo, la unión de estos tipos de parámetros logran una mejora en el rendimiento de la red en comparación con los dos tipos de parámetros por si solos. Lo mismo ocurre con las pruebas de Tipo B (Filtro r), donde si bien la diferencia entre los tipos de parámetros son menores, se mantienen los mejores resultados en la prueba con parámetros basados en color y perfiles ajustados. También se puede ver que la unión de los dos tipos de parámetros nos brinda un mejor rendimiento de la red, al igual que en las pruebas de Tipo A.

Analizando las pruebas de Tipo C, éstas se deben ver por separado y compararlas con respecto a la prueba 6 de Tipo B. Teniendo en cuenta esto último, la prueba 7 obtuvo una diferencia de un $0,5\,\%$ en el porcentaje de éxito con respecto a la prueba 6. Si bien es un porcentaje pequeño, podemos decir que la disminución del set de entrenamiento, debido a la selección de objetos brillantes, contribuye a esta diferencia, pero esto es algo que debiese comprobarse con la realización de otras pruebas adicionales, que en este caso, no es el objetivo del estudio. También podemos decir que la selección de objetos brillantes mantiene en cierta medida los resultados obtenidos en la prueba con los demás objetos con un filtro $r > 17$, por lo que esta selección no aporta a la mejora de rendimiento de la red.

Por otro lado, un resultado interesante tuvo que ver con la prueba 8. El objetivo de esta prueba era verificar si, eliminando el parámetro *texture*, se obtenía un resultado similar o no. En este caso, por lo visto en las tablas 3 y 4, los valores de rendimiento de la redes, en ambas pruebas, 6 y 7, son similares, variando los resultados solamente en la cuarta cifra significativa del MSE. Dado lo revisado en esta comparación, se puede decir que el parámetro *texture*, no contribuye a la clasificación que ocupa parámetros asociados al filtro r, por lo que se puede prescindir de él.

7. Conclusiones

El trabajo efectuado en la memoria permitió profundizar en el estudio teórico y práctico de las redes neuronales artificiales, además de llevar ese conocimiento al trabajo en el ámbito astronómico. Por otra parte, el conocer acerca del trabajo de los astrónomos ha resultado ser una experiencia alucinante que motiva a profundizar más en su campo.

Con respecto al trabajo, los resultados de las pruebas realizadas nos han otorgado conclusiones interesantes acerca del uso de ciertos parámetros, como el color, los perfiles ajustados de brillo, la forma y la textura, para el entrenamiento de la redes neuronales. Una de las conclusiones vista en el sub-capítulo de análisis de resultados, es acerca de los parámetros asociados al filtro r, y es que estos se comportan de mejor manera que los parámetros asociados al filtro u. Por otro lado, el rendimiento de las redes, que ocupan parámetros basados en color y perfiles ajustados de brillo, es mejor que el rendimiento de las que usan los parámetros de forma y textura. Otro resultado, y no menos importante, es que la variación de volumen del set de entrada no influye demasiado en el rendimiento global de la red que ocupa parámetros asociados al filtro r, lo cual se puede ver en la prueba con el set de datos de objetos brillantes. Por último, uno de los resultados más interesantes, es el visto en la prueba en cual se eliminaba el parámetro *texture*, que mide la gama de fluctuaciones en el brillo de la superficie de los objetos astronómicos. Este parámetro no aporta mayor información, a la red neuronal, que ayude a clasificar los objetos como galaxias de algún tipo, por lo que se puede prescindir de él para obtener la clasificación de galaxias ocupando parámetros asociados al filtro r. La eliminación del parámetro *texture* asociado a los demás filtros, se deja para una futura investigación.

Estos resultados obtenidos son importantes en términos de la clasificación de galaxias, ya que nos garantizan que ocupando un filtro fotométrico adecuado, del sistema *ugriz* de SDSS, y eligiendo los parámetros de entrada presentados en este trabajo, podemos obtener una buena clasificación de este tipo de objetos. A su vez, nos pone en un buen camino para desarrollar una máquina de aprendizaje que permita distinguir fusiones de galaxias tomando como base el trabajo realizado, lo cual era uno de los principales objetivos de esta memoria.

8. Trabajos Futuros

Como se demostró, las redes neuronales artificiales poseen un gran potencial para clasificar objetos astronómicos, como las galaxias. Esta ventaja se podría ocupar para otro tipo de clasificación o detección dentro del campo de la astronomía. Por ejemplo la identificación de fusiones de galaxias asoma como un gran desafío, ya que para los astrónomos, detectar esta mezcla de galaxias ha sido de gran dificultad. Se sabe que se han hecho algunos avances, como lo mostrado en Darg et al.[3][4][5], con respecto a fusiones y multi-fusiones de galaxias, lo cual invita a investigar sobre la posible ayuda que puedan ofrecer las máquinas de aprendizaje en la detección automática de estos objetos.

En el ámbito del tema de memoria, un posible trabajo sería mostrar el efecto de la elección de un umbral distinto en la clasificación de GZ1, lo cual podría generar importantes cambios en el rendimiento de la red neuronal. Esto porque, es decisión del investigador elegir el umbral adecuado para realizar sus pruebas, lo que podría provocar distintos rendimientos en la clasificación dependiendo del umbral elegido. Entonces, un posible trabajo sería cambiar el umbral en base a ciertos criterios, como lo visto en Banerji et al.[2]. Otra de las posibles propuestas sería repetir este experimento utilizando el catálogo del proyecto Galaxy Zoo 2, lo cual permitiría añadir otro tipo de parámetros que podrían mejorar la clasificación. Por otro lado, el procesamiento de las imágenes de galaxias sería otro de los trabajos a efectuar en el futuro, ya que se podría analizar y elegir otra clase de parámetros distintos a los ocupados en este trabajo. Por otro lado, se podrían realizar las mismas pruebas de este trabajo, pero considerando el algoritmo de entrenamiento de Levenberg-Marquardt. Ésto, para asegurar que converja la red neuronal en un tiempo menor al ocupado en las pruebas de este trabajo.

Referencias

[1] Bamford, S., Nichol, R., et al., *"Galaxy Zoo: the dependence of morphology and colour on environment"*. Monthly Notices of the Royal Astronomical Society (MNRAS), Vol. 393, pag. 1324-1352, 2009.

[2] Banerji, M., Lahav, O., et al., *"Galaxy Zoo: Reproducing galaxy morphologies via machine learning"*. Monthly Notices of the Royal Astronomical Society (MNRAS),Vol. 406, pag. 342-353, 2010.

[3] Darg, D., Kaviraj, S., Lintott, C., et al., *"Galaxy Zoo: the fraction of merging galaxies in the SDSS and their morphologies"*. Monthly Notices of the Royal Astronomical Society (MNRAS),Vol. 401, pag. 1043-1056, 2010.

[4] Darg, D., Kaviraj, S., Lintott, C., et al., *"Galaxy Zoo: the properties of merging galaxies in the nearby Universe-local environments, colours, masses, star formation rates and AGN activity"*. Monthly Notices of the Royal Astronomical Society (MNRAS),Vol. 401, pag. 1552-1563, 2010.

[5] Darg, D., Kaviraj, S., Lintott, C., et al., *"Galaxy Zoo: multimergers and the Millennium Simulation"*. Monthly Notices of the Royal Astronomical Society (MNRAS),Vol. 416, pag. 1745-1755, 2011.

[6] Fortson, L., Masters, K., et al., *"Galaxy Zoo: Morphological Classification and Citizen Science"*. Advances in Machine Learning and Data Mining for Astronomy, CRC Press, p. 213-236, 2011.

[7] Kennicutt, R., *"A Spectrophotometric Atlas of Galaxies"*. The Astrophysical Journal supplements series, Vol. 79, 1992.

[8] Lintott, C., Schawinski, K., et al., *"Galaxy Zoo 1 : Data Release of Morphological Classifications for nearly 900,000 galaxies"*. Monthly Notices of the Royal Astronomical Society (MNRAS), Vol. 410, pag. 166-178, 2010.

[9] Visa, S., Ramsay, B. et al., *"Confusion Matrix-based Feature Selection"*. Midwest Artificial Intelligence and Cognitive Science Conference, USA, 2011.