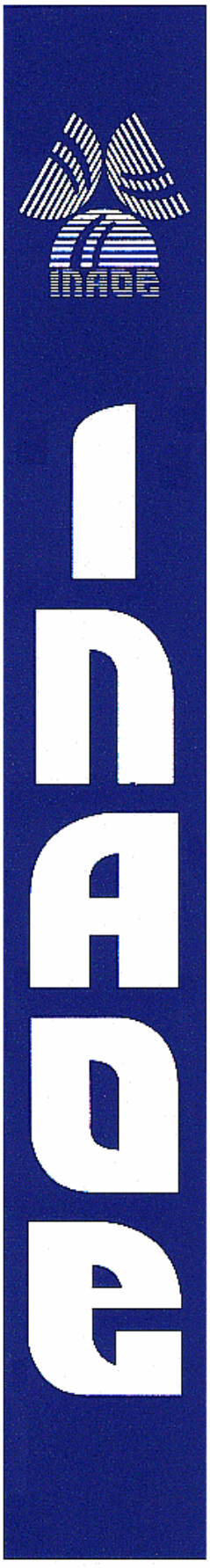

Modelo Jerárquico para la Clasificación de Galaxias

Por

Ing. Maribel Angélica Marín Castro

Tesis sometida como requisito parcial para obtener el grado de

MAESTRO EN CIENCIAS EN LA ESPECIALIDAD DE CIENCIAS COMPUTACIONALES

en el

Instituto Nacional de Astrofísica, Óptica y Electrónica.

Supervisada por:

Dr. Luis Enrique Sucar Succar.
Investigador del INAOE

Dr. Jesús A. González Bernal.
Investigador del INAOE

Dra. Raquel Díaz Hernández
Investigador del INAOE

Agradecimientos

A Dios por nunca abandonarme y estar conmigo en cada paso que doy.

A mis asesores los Doctores Luis Enrique Sucar Succar, Jesús González Bernal y Raquel Díaz Hernández por su gran apoyo y motivación, por haberme compartido su tiempo y conocimiento y por dirigirme en el desarrollo de este trabajo de tesis. Así también quiero agradecer a los Doctores Ariel Carrasco, Carlos Reyes y Enrique Muñoz de Cote por el tiempo dedicado a la revisión de esta tesis y por sus valiosas sugerencias.

A mis padres Miguel y Julia por haberme apoyado en todo momento, por sus consejos, por mostrarme que siempre hay que buscar el ser alguien mejor en la vida y por el amor que siempre me han demostrado. También agradezco a mis hermanos Heidy y Miguel Angel por estar conmigo y apoyarme siempre.

A mi novio Julio, por no rendirse conmigo al buscar la forma de mostrarme la persona que realmente soy, por todos sus ánimos y sobre todo por ser siempre mi consuelo y apoyo.

A mis amigos del INAOE quienes me han apoyado en situaciones difíciles y con quienes he compartido momentos de nerviosismo: Juan Manuel Cabrera, Lucas Pacheco, Alejandro Torres, Alejandro Rosales, Aaron Rocha, Adrian Leal, Majandy (Alejandra Menéndez) y Marisol Flores quien simpre nos motivo y compartio su conocimiento con la treceava generacion.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por otorgarme una beca para llevar acabo mis estudios de maestría. Al Instituto Nacional de Astrofísica Óptica y Electrónica (INAOE) por la formación académica y los apoyos y servicios que me brindó.

"A mi familia por su gran apoyo y porque simpre han estado a mi lado en los momentos difíciles."

Resumen

El Instituto Nacional de Astrofísica, Óptica y Electrónica posee alrededor de 50 años de investigación astronómica contenidos en su colección de placas astronómicas tomadas en la Cámara Schmidt. A pesar de contar con esta gran colección no se han analizado todos los objetos estelares que se encuentran en ellas. Un caso particular de estos objetos estelares son las galaxias, por lo que en este trabajo de tesis se realizó la clasificación morfológica de las mismas. Se propone un nuevo método para la clasificación morfológica de galaxias basado en un enfoque jerárquico. El método propuesto inicia desde el procesamiento de la placa, para identificar los objetos estelares contenidos en la misma. Una vez identificados los objetos, se lleva a cabo una primera clasificación para discernir entre estrellas y galaxias. Después de separar estrellas de galaxias, entonces nos enfocamos en los objetos de interés, en este caso nos referimos a las galaxias. A continuación, se toma la imagen de la galaxia y se lleva a cabo la separación el núcleo del resto de la galaxia, para así poder hacer un análisis de ambos. Se diseño una taxonomía basada en la clasificación de Hubble para llevar a cabo una clasificación jerárquica utilizando una técnica de clasificación jerárquica multidimensional. Debido a que existe una mayor cantidad de galaxias espirales en comparación a las galaxias de otros tipos, fue neceario tratar el problema de desbalance de clases y para ello se realizaron dos tareas. La primera fue crear ejemplos artificiales (imágenes) aplicando transformaciones geométricas a las imágenes originales y la segunda fue el utilizar una técnica de sobre-muestreo. El método propuesto fue probado con una base de datos creada en este mismo proyecto de tesis. Los resultados mostraron que la clasificación jerárquica tiene un desempeño mayor que la clasificación plana, y además que al agregar datos mediante sobre-muestreo y ejemplos artificiales también se mejora el desempeño de clasificación.

Abstract

The National Institute for Astrophysics, Optics, and Electronics owns an astronomical plates collection that surveys about 50 years of research. These astronomical plates were captured with a Schmidt Camera, and although they were obtained long ago, not all the stellar objects found in there have been analyzed. A particular case of these stellar objects consists of galaxies, for which this research work focus in their morphological classification. In order to perform this classification, we require the use of images processing and classification (such as hierarchical classification) techniques. The state of the art makes reference to diverse alternatives to solve the galaxies classification problem. Some of the more relevant works use classification ensembles and consider that this domain suffers the class imbalance problem. In order to perform a better analysis of galaxies, some researchers use a technique that separates the galaxy nucleus from the rest of its body. However, in all cases, the main problem worseness as the number of types of galaxies to classify increases. This considerably decreases the classification accuracy. In this work, we propose a new method for the morphological galaxies classification based on a hierarchical technique. The proposed method starts with the processing of a plate in order to identify the stellar objects found in it. Once these objects have been identified, we perform an initial classification to separate between stars and galaxies. After separating stars from galaxies, we focus in our objects of interest; in this case, we refer to galaxies. Next, we take the galaxy image and separate its nucleus from the rest of the galaxy so that we can consider both (separately) in the analysis. We designed a hierarchy based on the Hubble's classification in order to perform a hierarchical classification using a multi-dimensional hierarchical classification technique. In order to deal with the class imbalance problem we performed two tasks. In the first one, we create artificial examples (images) applying geometric transformations to the original images. In the second one, we use over sampling technique. The proposed method was tested with the database created for this thesis project. The results found show that the hierarchical classification method has a higher performance than the flat classification technique.

Índice general

Índice de Tablas

Índice de Figuras

Capítulo 1

Introducción

La fotografía astronómica permite registrar instantes de la evolución estelar. Hoy en día, dentro de la astronomía, la astro-fotografía sigue siendo una herramienta fundamental de apoyo para cualquier astrónomo, a pesar de estar en una época en la cual la mayoría de las investigaciones se llevan a cabo por medios electrónicos.

En la actualidad, en el área de astrofísica se ha presentado un gran desarrollo gracias a la aparición de nuevos detectores cada vez más sensibles y al aumento en la capacidad de recolección y almacenamiento de datos, por lo que se cuenta con una gran cantidad de información de los objetos estelares. Sin embargo las técnicas tradicionales de reducción, clasificación y análisis de las observaciones estelares no permiten el manejo eficiente de tal cantidad de información. Debido a esto, el estudio de estos objetos debe hacerse de manera particular siguiendo una vertiente de análisis dependiendo de los tipos de objetos estelares. La clasificación de galaxias es muy importante por dos razones. La primera es que produce grandes catálogos para programas estadísticos y de observaciones; y la segunda es que permite descubrir la física subyacente como se explica en el trabajo de Lahav [Lahav, 1996].

Existen dos formas de abordar la clasificación de galaxias: morfológica y espectral. La clasificación basada en la morfología de la galaxia describe la apariencia de la misma, mientras que aquella que se basa en el espectro de la galaxia considera medidas de composición estelar. Se debe entender que una es complemento de la otra y no verlas

como algo separado. Una de las primeras clasificaciones de galaxias basadas en su morfología fue dada por Edwin Hubble en 1926. En esta clasificación se dividen las galaxias en tres tipos principales que son: elípticas, espirales e irregulares, pero con el transcurso del tiempo esta clasificación se ha ido refinando. Algunas de las técnicas utilizadas para dar solución al problema de clasificación de galaxias son las redes neuronales, árboles de decisión, ensambles de clasificadores y métodos basados en instancias, como se muestra en el trabajo de De la Calleja y colaboradores [De la Calleja et al. 2004].

1.1. Motivación

El Antiguo Observatorio Astrofísico Nacional de Tonantzintla en el estado de Puebla, hoy INAOE, fue uno de los lugares en el mundo donde se hicieron estudios profundos con placas astronómicas, realizando un gran muestreo del cielo en un periodo de 50 años, como se comenta en el trabajo de [Diaz, 2005]. Dado que el INAOE cuenta con una colección importante de placas astronómicas, formada por 15,456 placas, las cuales se encuentran distribuidas en placas espectrales, placas con imagen directa y placas de tres imágenes (directa), como se muestra en la Figura 1.1, surgió la necesidad de preservar y estudiar dichas placas de forma automática. Las placas astronómicas, a pesar de su antigüedad, aún contienen información de objetos que no han sido estudiados en detalle, por lo que no se conoce mucho de ellos. El desarrollo de técnicas para la extracción y la clasificación automática de información de estas placas es de suma importancia, pues de esta forma se puede saber más sobre objetos casi desconocidos.

1.2. Problemática

Uno de los principales problemas en la clasificación de galaxias en placas astronómicas es el hecho de que en las placas astronómicas se pueden confundir los objetos estelares, ya sea entre ellos mismos o con defectos de la placa. Sumado a lo anterior, existen galaxias que se encuentran más lejanas y por lo tanto no se logra apreciar de manera óptima la forma de éstas provocando problemas en su clasificación. También se encuentra presente el problema de las clases desbalanceadas, ya que se puede obtener mayor

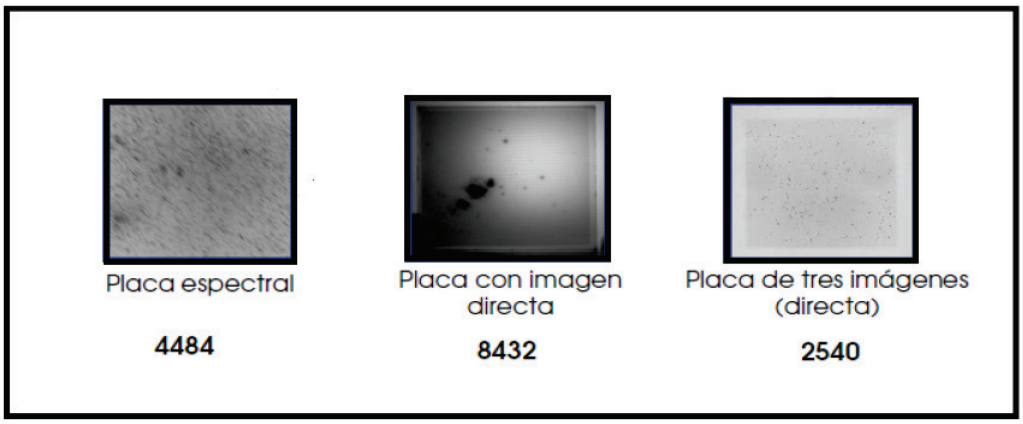

Figura 1.1: Colección de placas astronómicas del INAOE: ejemplo de placas astronómicas.

número de galaxias espirales y elípticas que de galaxias irregulares, por mencionar un ejemplo. Además, se presenta el problema del traslape de galaxias, esto es, que una galaxia se encuentre detrás de otro cuerpo estelar y esto no permite observar correctamente su morfología.

Los principales problemas que se presentan al realizar el proceso de clasificación de las galaxias en placas astronómicas son:

- Las imágenes astronómicas contienen una gran cantidad de objetos estelares, de los cuales para lograr los objetivos de este trabajo, se necesita identificar y clasificar únicamente las galaxias, pero muchos de ellos pueden confundirse entre sí, complicando el proceso de identificación y clasificación de las galaxias como se puede observar en la Figura 1.2.

- Debido a problemas en la emulsión de la placa fotográfica, las imágenes presentan defectos que pueden ser confundidos con objetos estelares como se muestra en la Figura 1.3

- Existen galaxias que se encuentran demasiado retiradas y por lo tanto no se puede apreciar bien su morfología.

- En ocasiones existe el traslape de objetos estelares, es decir, cuando un objeto se encuentra obstruyendo la visibilidad de otro, como se puede observar en la Figura

Figura 1.2: Ejemplo de placa astronómica. Como se puede observar se encuentran tres galaxias en esta placa, la primera se ubica en la parte superior izquierda, dos se encuentran en la parte inferior derecha y el resto de los presentes objetos son estrellas.

1.4.

- Las estrellas más brillantes producen saturación, lo cual genera ruido en el proceso de segmentación de la imagen como se muestra en la Figura 1.5.

1.3. Objetivos

1.3.1. Objetivo General

Diseñar e implementar un algoritmo con un enfoque jerárquico, capaz de segmentar y clasificar galaxias contenidas en placas astronómicas, de acuerdo a su morfología.

1.3.2. Objetivos Específicos

- Desarrollar un método de segmentación y extracción de características de galaxias para imágenes astronómicas.
- Diseñar un esquema de clasificación jerárquico orientado a la clasificación de galaxias.
- Desarrollar un método para tratar el problema de desbalance de clases.

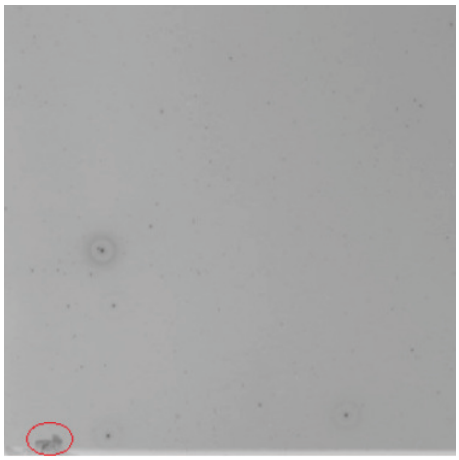

Figura 1.3: Ejemplo de placa astronómica. Se puede observar un defecto de emulsión en el lado inferior izquierdo.

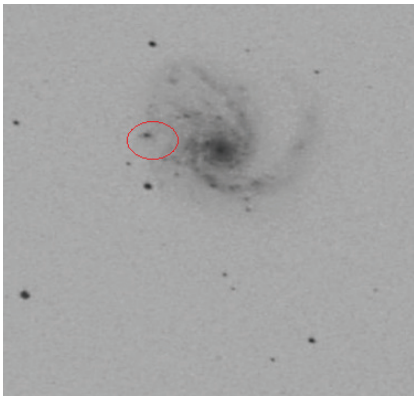

Figura 1.4: Se ilustra un ejemplo del problema de traslape.

1.4. Solución Propuesta

En esta sección se describe la metodología que se seguirá para alcanzar los objetivos planteados anteriormente. Esta metodología se ilustra en la Figura 1.6. A continuación se describe cada uno de los pasos a seguir en la metodología.

1. Seleccionar y digitalizar las placas astronómicas con mayores cúmulos de galaxias.

2. Segmentar las galaxias de las placas astronómicas y después llevar a cabo la separación del núcleo con el resto del cuerpo de la galaxia.

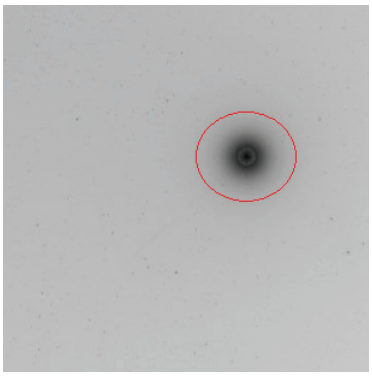

Figura 1.5: Ejemplo de placa astronómica. Se ilustra el problema de una estrella saturada.

3. Extraer las características principales de las imágenes de las galaxias segmentadas.

4. Analizar e implementar la solución para tratar el problema de las clases desbalanceadas.

5. Diseñar e implementar un algoritmo de clasificación jerárquica.

6. Elaborar los experimentos con la base de datos obtenida de las placas astronómicas de INAOE.

7. Realizar la evaluación del método propuesto.

En este trabajo se propusó un método que permite en primer lugar segmentar y extraer características de los objetos contenidos en las placas para así separar estrellas de galaxias. En esta primera parte se obtuvo el 91.66 % de precisión en la clasificación, utilizando el clasificador *Random Forest* y momentos geométricos, de cada uno de los objetos, como atributos para la clasificación. Después se llevó a cabo la separación de núcleo de la galaxia del resto del cuerpo de la misma, utilizando de igual manera los momentos geométricos y el radio de Petrosian. Una vez segmentada la imagen de la galaxia se realizó la extracción de características y se llevó a cabo una clasificación plana. En esta clasificación el clasificador que obtuvo mejores resultados fue *Random Forest* con el 64.17 % para cuatro tipos de galaxias (E, S, S0, Irr) y 39.44 % para nueve tipos (E, Sa, Sb, Sc, SBa, SBb, SBc, S0, Irr), utilizando ejemplos artificiales y *Resampling*. Por último, se realizó la clasificación jerárquica, para esta clasificación se utilizó la técnica de

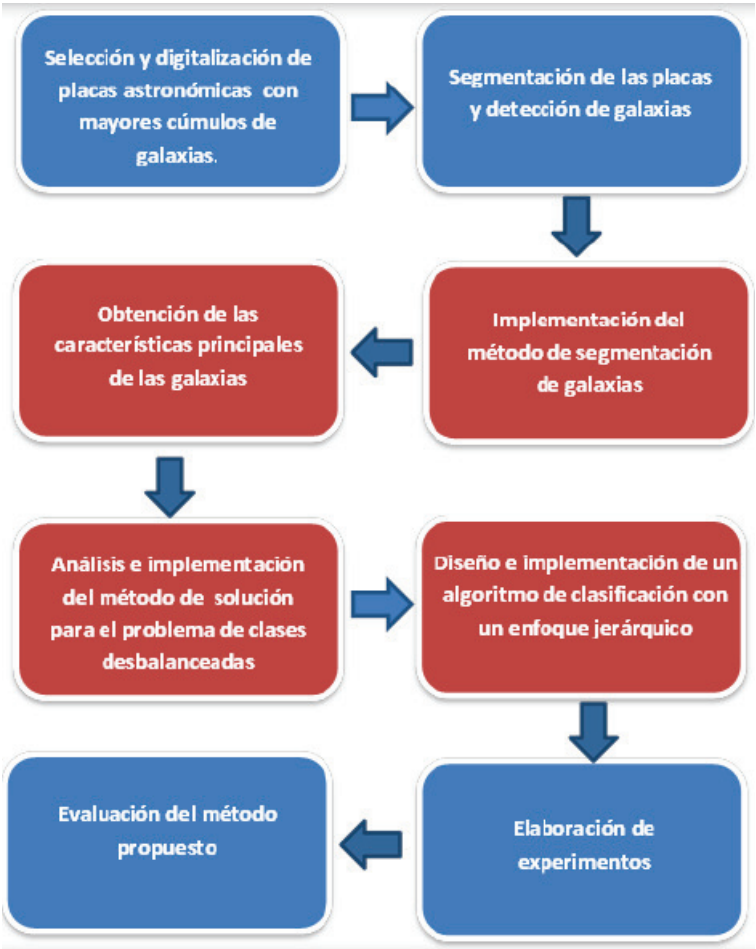

Figura 1.6: Diagrama de la Metodología. Se ilustran los pasos a seguir para alcanzar el objetivo general de este trabajo. Los bloques que se encuentran de color rojo son la parte principal de este proyecto de tesis.

clasificación jerárquica multidimensional con la variante de producto de probabilidades y nuevamente el clasificador con mejores resultados fue *Random Forest* con el 53.57 % para nueve tipos de galaxias.

1.5. Organización del Documento

En el siguiente capítulo se presentan conceptos del aprendizaje supervisado. Se expone la estructura y funcionalidad de algunos algoritmos como *Naive Bayes* y *Random Forest*, así como algunas técnicas de clasificación jerárquica. También se presentan algunos de los mecanismos para evaluar el desempeño de un clasificador.

En el capítulo 3 se presentan conceptos básicos de astronomía que son utilizados en la clasificación de galaxias.

En el capítulo 4 se describe el proceso de análisis de imágenes, así como las dos tareas que se encuentran relacionadas con este proceso, la segmentación de imágenes y la extracción de características.

En el capítulo 5 se describe el diseño e implementación de la estrategia de clasificación de galaxias propuesta.

En el capítulo 6 se muestran los resultados obtenidos de las pruebas realizadas con la base de datos de imágenes de la colección de placas del INAOE.

Por último, en el capítulo 7, se presentan las conclusiones de este trabajo de tesis y alguna direcciones del trabajo futuro.

Capítulo 2

Clasificación Automática

El Aprendizaje Automático es considerado como la disciplina que se encarga de estudiar el cómo construir sistemas computacionales que realicen cada vez mejor una tarea mediante la experiencia. Es decir, se considera que un programa es capaz de aprender a desarrollar cierta tarea T si después de proporcionarle la experiencia E es capaz de desempeñar correctamente esa tarea ante nuevas situaciones. A su vez, este desempeño es evaluado usando una métrica de calidad M. Por lo tanto, un problema de aprendizaje bien definido requiere que T, E y M estén bien especificados [Mitchell, 1997].

En este capítulo nos enfocaremos a algunas de las tecnicas de aprendizaje automático como son la clasificación y el sobremuestreo. Estas técnicas serán descritas con mayor detalle en las siguientes secciones.

2.1. Clasificación

El término de *clasificación* se refiere a la descripción general de un objeto como perteneciente a una clase natural de objetos semejantes [Ullman, 2001]. El objetivo de los algoritmos de clasificación es inducir un modelo para predecir la clase que identifica un objeto dados los valores de los atributos o características que lo representan. Enseguida se describen los algoritmos utilizados en esta tesis para clasificar los distintos tipos de galaxias.

2.1.1. Algoritmos de Clasificación Supervisada

Los algoritmos de clasificación supervisada se utilizan en problemas en los cuales se conoce *a priori* el número de clases y algunos representantes de cada clase. Es decir, estos algoritmos operan usualmente sobre la información suministrada por un conjunto de muestras, patrones, ejemplos o prototipos de entrenamiento que son asumidos como representantes de las clases, y los mismos poseen una etiqueta de clase correcta. A este conjunto de prototipos correctamente etiquetados se le llama conjunto de entrenamiento (TS, *training set*), y es el conocimiento empleado para generar un modelo de clasificación para nuevas muestras. Su objetivo es determinar cuál es la clase, de las que ya se tiene conocimiento, a la que debe pertenecer una nueva muestra.

Naive Bayes

El clasificador Bayesiano, es aquel que se encuentra basado en la regla de Bayes [Michie et al., 1994], este obtiene la probabilidad de cada una de las clases C_i, como el producto de la probabilidad a *priori* por la probabilidad condicional de los atributos E=$\{E_1$,E_2,...,$E_n\}$ dada la clase. Es decir:

$$P(C_i|E) = \frac{P(C_i)P(E|C_i)}{P(E)} \tag{2.1}$$

Donde $P(C_i)$ es la probabilidad de cada clase C_i, $P(E)$ es la probabilidad de los atributos y $P(E|C_i)$ es la probabilidad condicional de los atributos dada la clase.

El clasificador *Naive Bayes* asume que todos los atributos son independientes entre si dada la clase. Por lo tanto, la probabilidad se puede obtener por el producto de las probabilidades condicionales individuales de cada atributo dada la clase, esto se muestra en la siguiente ecuación:

$$P(C_i|E) = \frac{P(C_i)P(E_1|C_i)P(E_2|C_i)...P(E_n|C_i)}{P(E)} \tag{2.2}$$

Una vez calculadas las probabilidades individuales de cada atributo, el clasificador *Naive Bayes* regresa las probabilidades de las clases. El clasificador *Naive Bayes* tiene

la ventaja de ser sencillo de construir y entender, además de ser rápido para realizar inducciones.

Random Forest

Random Forest (RF) es un clasificador basado en ensambles de árboles de decisión sin poda. Para llevar a cabo la clasificación de una nueva instancia, el vector de características de dicha instancia es colocado en cada uno de los árboles de decisión. Entonces, cada árbol genera una clasificación, es decir cada árbol vota por una clase. A través de un sistema de votación se decide la clase, ya que toma la clase con un mayor número de votos. Los árboles de decisión son construídos con una muestra aleatoria de los atributos [Breiman, 2001]. El algoritmo para la construcción de un clasificador RF se muestra a continuación:

Algoritmo de Random Forest

Requiere : **IDT**(Algoritmo para construir un árbol de decisión),
T(número de iteraciones),
S(conjunto de entrenamiento),
μ(tamaño de la muestra),
N(número de atributos usados en cada nodo)

Garantizar: M_t;$t = 1, 2, ..., T$

para $t \leftarrow 1$ **hasta** T **hacer**
$S_t \leftarrow$ muestra con μ instancias de S con remplazo
Construir el clasificador M_t usando **IDT**(**N**) en S_t
$t++$
fin para

Algunas de las ventajas de **Random Forest** son:

- Se ejecuta de manera eficiente en grandes bases de datos.

- Ofrece una estimación de las variables que son importantes en la clasificación.
- Tiene un método eficaz para la estimación de datos faltantes y mantiene la precisión aun con una gran proporción de datos faltantes.
- Los árboles generados se pueden guardar para uso futuro con otros datos.

2.2. Clasificación Jerárquica

La clasificación jerárquica puede ser vista como un problema particular de la clasificación estructurada [Astikainen et al., 2008], donde la salida del algoritmo de clasificación está dada por una taxonomía. En este caso, la jerarquía tiene una estructura organizada por niveles y ramas. En cada rama se distribuyen los elementos de la jerarquía, de lo general a lo particular. Por ejemplo se sabe que un conjunto de objetos son galaxias, pero estos a su vez pueden dividirse en espirales, elípticas e irregulares, así mismo las galaxias espirales se pueden dividir en espirales barras y espirales normares y así sucesivamente. La taxonomía en una clasificación puede tener una estructura de grafo acíclico dirigido (DAG) o bien un árbol.

Existen dos formas básicas para llevar a cabo la exploración de la jerarquía, es decir el recorrido que se hace sobre de cada uno de los nodos de la jerarquía. La primera es con clasificadores locales (*top-down*) y la segunda con clasificadores globales (*big-bang*). Pero el primer enfoque no es un enfoque completamente de clasificación jerárquica por sí mismo, sino más bien un método para evitar las inconsistencias en la predicción de la clase a diferentes niveles, durante la fase de entrenamiento. En el caso del segundo enfoque, los clasificadores son entrenados para considerar toda la jerarquía en un solo paso.

Los principales métodos para realizar la clasificación jerárquica son:

- *Clasificación Jerárquica Plana*. Este método de clasificación jerárquica es el más simple [Xiao et al.,2007] pues consiste en ignorar completamente la jerarquía de clases, por lo que la predicción de las clases se lleva a cabo sólo en los nodos hoja.

Este enfoque se comporta como un algoritmo de clasificación tradicional durante el entrenamiento y pruebas. Sin embargo, proporciona una solución indirecta al problema de la clasificación jerárquica, ya que, cuando una clase de la hoja se le asigna a un ejemplo, se puede considerar que todas sus clases antecesoras son también implícitamente asignadas a esa instancia. Sin embargo, este enfoque tiene el inconveniente de que existe un solo clasificador para discernir entre un gran número de clases.

- *Clasificación Jerárquica Local (Top-Down)*. En este caso la jerarquía se tiene en cuenta mediante el uso de la información local [Koller et al.,1997], la cual puede ser empleada de diferentes maneras. Estos enfoques, por lo tanto, se pueden agrupar en función de cómo utilizan esta información local y cómo construyen los clasificadores que lo rodean. Por lo tanto existen tres formas de usar la información local: un clasificador local por nodo (LCN), un clasificador local por nodo padre (LCPN) y un clasificador local por nivel (LCL).

  - Clasificador Local por Nodo. Este enfoque es el más utilizado en la literatura, y consiste en la creación de un clasificador binario para cada nodo de la jerarquía (excepto el nodo raíz).
  - Clasificador Local por Nodo Padre. En este caso se refiere a que los clasificadores sólo se encuentran en los nodos padres de la jerarquía y no se crea ningún clasificador en los nodos hoja.
  - Clasificador Local por Nivel. En este enfoque se construye un clasificador multi-clase por cada nivel de la jerarquía.

  Una desventaja en este tipo de enfoque es que un error en el clasificador superior será propagado a los demás niveles.

- *Clasificación Jerárquica Global (Big-Bang)*. En este caso se crea un modelo global de clasificación [Freitas et al.,2007]. Tiene la ventaja de que el tamaño total del modelo de clasificación global es generalmente mucho menor, en comparación con el tamaño total de todos los modelos locales en cualquiera de los enfoques de clasificación local. Su desventaja es que al aumentar o disminuir el número de clases se debe rehacer el modelo.

2.3. Clasificación Jerárquica Multidimensional

Como se mencionó en la sección anterior, la clasificación jerárquica es una variante de la clasificación multidimensional, con la diferencia de que las clases se encuentran organizadas dentro de una jerarquía. En este caso se puede considerar a la clasificación jerárquica multidimensional [Hernandez,2012] como un enfoque de clasificación jerárquico que utiliza uno de tres metodos para la predicción de las clases, los cuales se explicaran en esta sección.

La clasificación jerárquica mutidimensinal se inicia con la construcción de un clasificador multi-clase (aquel que predice una clase de entre varias clases) por cada nodo primario en la jerarquía en la fase de entrenamiento. En la fase de clasificación, a diferencia de un enfoque tradicional *top-down*, todos los clasificadores locales se aplican simultáneamente para cada una de las instancias, así que cada clasificador local obtiene una probabilidad para cada clase.

Se asume una taxonomía *T* de tipo árbol, con *t* nodos, donde cada nodo representa una clase. Hay *c* nodos no-hoja y *l* nodos hoja, de manera que *t*=*c*+*l*. Cada nodo no-hoja c_i tiene ns_i(nodos hijo) que representan las clases directas de la clase c_i. Se asume que hay *m* atributos para cada clase de tal manera que el mismo conjunto de atributos es considerado por todas las clases, es decir, que todas las clases cuentan con el mismo número de atributos. A continuación se presenta el algoritmo de este método de clasificación jeárquica.

Algoritmo de Clasificación Jerárquica Multidimensional

Entrenamiento: Dada una base de datos de n puntos de datos (x_1, j_1) ,..., $(x_n; j_n)$ donde x_i son los m atributos y j_i la clase de acuerdo a una taxonomía T.

1. Dividir la base de datos de acuerdo a las subclases (hijo) de cada nodo padre c_i.
2. Entrenar un clasificador multi-clase de cada nodo padre para clasificar sus nodos hijo.

Clasificación: Dada una instancia **x**:

1. Clasificar **x** en todos los clasificadores locales c.
2. Combinar los resultados de todos los clasificadores para producir el camino más probable (de la raíz a la hoja).

Existen tres formas diferentes de llevar a cabo esa combinación de los resultados de los clasificadores locales, las cuales son: Ordenamiento Descendente de Probabilidades (**ODP**), Producto de Probabilidades (**PP**) y Suma de Probabilidades (**SP**). Las cuales se explicarán a continuación.

- **Ordenamiento Descendente de Probabilidades (ODP)**. Las clases predichas por los clasificadores locales son ordenadas de forma descendente según sus probabilidades. De acuerdo con este orden se busca el primer subconjunto consistente de clases; es decir, una ruta desde la raíz hasta una de las hojas, entonces este conjunto se devuelve como la predicción global.
- **Producto de Probabilidades(PP)**. Este método multiplica las probabilidades de la clase más probable para todos los nodos de cada camino desde la raíz hasta una hoja en la jerarquía. La predicción global será el conjunto de clases en el camino con el mayor producto.

- **Suma de probabilidades(SP)**. En este caso se suman las probabilidades de la clase más probable de cada uno de los nodos locales de cada camino de la jerarquía. Por lo tanto, la predicción global será el camino con la suma más alta.

Un ejemplo de este método de clasificación jerárquica se puede observar en la Figura 2.1.

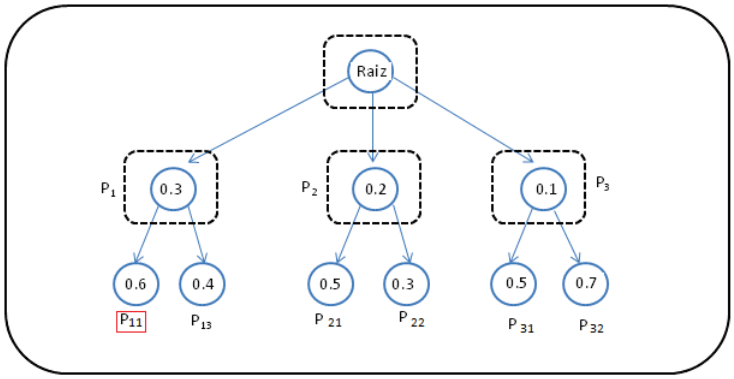

Figura 2.1: Clasificación Jerárquica Multidimensional. En la imagen cada nodo (círculo) representa a cada una de las clases de la jerarquía, las líneas punteadas representan a los clasificadores y los números dentro de los nodos son las probabilidades de cada clase. Por lo tanto en este ejemplo se tiene como la clase predicha a P_1 y P_{11}, pues **PP**=0.18, **SP**=0.9. Y para **ODP** se tienen las probabilidades ordenadas de la siguiente manera $P_{32}, \mathbf{P_{11}}, P_{21}, P_{31}, P_{13}, P_{22}, \mathbf{P_1}, P_2, P_3$, por lo tanto **ODP**=$P_1, P_{11}$.

2.4. Evaluación del clasificador

Un factor importante en el área de la clasificación es la evaluación del desempeño del clasificador. La forma de llevar a cabo esa evaluación de la eficiencia de un clasificador es mediante la precisión predictiva del mismo. Cuando se introducen nuevos ejemplos a un clasificador, éste debe tomar la decisión sobre la clase que le asignará a cada uno de estos ejemplos. La clasificación incorrecta de un ejemplo se considera como un error del clasificador. La tasa de error se calcula como:

$$Tasa\ de\ error = \frac{Numero\ de\ errores}{Numero\ total\ de\ casos} \tag{2.3}$$

Existen otras medidas de evaluación del desempeño de un clasificador, las cuales se describen a continuación:

- *Verdaderos Positivos* (TP): Son aquellas instancias cuya hipótesis dice que deben ser positivas y en realidad son positivas.
- *Verdaderos Negativos* (TN): Son aquellas instancias cuya hipótesis dice que deben ser negativas y en realidad son negativas.
- *Falsos Positivos* (FP): Son aquellas instancias cuya hipótesis dice que deben ser positivas y en realidad son negativas.
- *Falso Negativo* (FN): Son aquellas instancias cuya hipótesis dice que deben ser negativas y en realidad es positivas.

A partir de las medidas mencionadas anteriormente surgen otras medidas de evaluación de un clasificador las cuales son:

- *Precision*: Es el porcentaje de predicciones positivas que son correctas, es decir, es la probabilidad de que una ejemplo x sea clasificada con la clase c, y que realmente pertenezca a esta clase. Y ésta se obtiene de la siguiente manera:

$$Precision = \frac{TP}{(TP + FP)} \tag{2.4}$$

- *Recall*: Es el porcentaje de verdaderos positivos predichos de entre todos los positivos, es decir, es la probabilidad de que si una instancia x pertenece a una clase c, el clasificador la clasifique correctamente.

$$Recall = \frac{TP}{(TP + FN)} \tag{2.5}$$

- *Accuracy*: Es el porcentaje de predicciones que son correctamente clasificadas

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \tag{2.6}$$

- *Especificidad*: Es el porcentaje de instancias negativas que fueron predichas como negativas.

$$Especificidad = \frac{TN}{(TN + FP)} \tag{2.7}$$

Existen diversas estrategias de validación de un sistema de aprendizaje a continuación se describirán dos de ellas.

Validación Simple

Este es el método de validación más sencillo [Sanchez,2005], ya que solo utiliza un conjunto de muestras para construir el modelo del clasificador. De entre la variedad de porcentajes utilizados, uno de los más frecuentes es tomar 2/3 del conjunto total de ejemplos para el proceso de aprendizaje y 1/3 para comprobar el error del clasificador.

Validación Cruzada

Esta técnica consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones [Kohavi et al.,1995]. Es decir, el conjunto de datos *D* se divide aleatoriamente en k particiones mutuamente exclusivas, *D1,D2,....Dk*, conteniendo cada una el mismo número de ejemplos aproximadamente. La validación cruzada se ejecuta k-veces, por ello esta técnica es llamada en muchos casos como validación cruzada con k particiones o *k-fold-cross validation*. Un valor comúnmente usado para k es 10. En cada evaluación se utiliza uno de los subconjuntos como conjunto de prueba, y se entrena el sistema con los $k - 1$ conjuntos restantes. Así, la precisión estimada de clasificación es la media de las k tasas obtenidas. Un ejemplo de esto se puede observar en la Figura 2.2 La ventaja de usar *k-fold-cross validation* es que todos los ejemplos en el conjunto de datos son eventualmente usados para entrenamiento y prueba. La estimación de la precisión de *k-fold-cross validation* es el número completo de clasificaciones correctas sobre el número de instancias en el conjunto de datos. La validación cruzada se realizan tantas iteraciones como muestras (N) tenga el conjunto de datos. De forma que para cada una de las N iteraciones se realiza un cálculo de error. El

resultado final lo obtenemos realizando la media aritmética de los N valores de errores obtenidos, como se muestra en la siguiente fórmula:

$$E = \frac{1}{N} \sum_{i=1}^{N} E_i \tag{2.8}$$

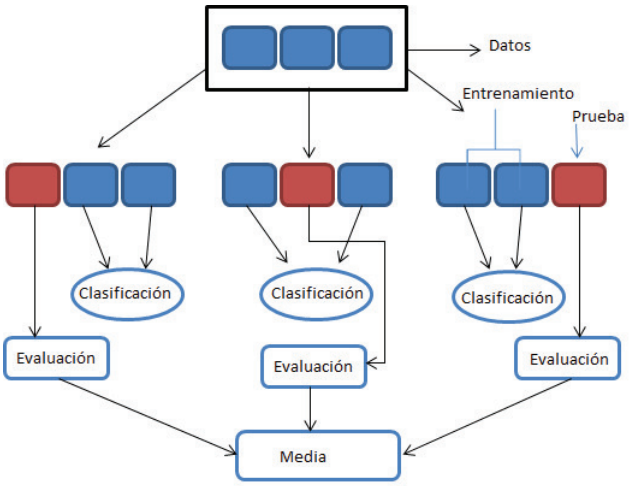

Figura 2.2: Validación Cruzada

2.5. Clases Desbalanceadas

Los algoritmos de aprendizaje, al ser utilizados en problemas reales, deben enfrentarse a diversos desafíos. Uno de estos desafíos es el problema de las clases desbalanceadas; es decir, puede haber una gran cantidad de ejemplos para una clase pero muy pocos ejemplos para la clase de interés. Este problema puede provocar malos resultados en la clasificación, por lo que se busca una forma de tratarlo.

Existen tres formas generales de tratar este problema [Morales et al.,2010], las cuales son:

- Mejores medidas para realizar la evaluación del desempeño de clasificadores ante clases desbalanceadas.

- Utilizar técnicas que sesgan los algoritmos para mejorar la clasificación de la clase minoritaria.
- Algoritmos que sesgan los datos para aumentar la cantidad de datos de la clase minoritaria, o disminuir los de la mayoritaria

En este trabajo de tesis nos enfocamos a los algoritmos que sesgan los datos. Éstos a su vez se dividen en (i) sobre muestreo (*over-sampling*) y (ii) submuestreo (*under-sampling*). En el primer caso se trata de generar más ejemplos de la clase minoritaria, mientras que en el segundo caso lo que se hace es eliminar ejemplos de la clase mayoritaria. Dos de los métodos más utilizados de sobre muestreo son **SMOTE** y **Resampling**, los cuales se describirán a continuación.

SMOTE

SMOTE (*Synthetic Minority Over-sampling Technique*), es un algoritmo de sobre-muestreo para ejemplos de la clase minoritaria. Este sobre-muestreo se realiza de la siguiente manera:

1. Se le proporciona el porcentaje de ejemplos a sobre-muestrear.
2. Calcular el número de ejemplos que tiene que generar.
3. Calcular los k vecinos más cercanos de los ejemplos de la clase minoritaria.
4. Por cada ejemplo de la clase minoritaria, elige aleatoriamente el vecino que utilizará para crear el nuevo ejemplo.
5. Para cada atributo del ejemplo, se calcula la diferencia entre el vector de atributos muestra y el del vecino elegido.
6. Se multiplica la diferencia anterior por un número aleatorio en el intervalo de 0 y 1.
7. Se suma el valor de la multiplicación al valor original de la muestra.
8. Por último, se regresa el conjunto de ejemplos sintéticos.

Resampling

El término *Resampling* hace referencia a aquellas técnicas empleadas en la teoría de probabilidades e inferencia estadística, las cuales a partir de datos observados, generan nuevas muestras simuladas de igual tamaño que la muestra original [Simon et al. 1991]; es decir, la simulación de los datos debe estar basada en algunos datos reales. Es decir, produce una submuestra aleatoria de un conjunto de datos utilizando el muestreo con reemplazo.

En el ámbito de la estadística, se denomina remuestreo a una variedad de métodos que permiten realizar algunas de las siguientes operaciones:

- Estimar la precisión de muestras estadísticas (mediana, variancia, percentil) mediante el uso de subconjuntos de datos disponibles (*jackknifing*) o tomando datos en forma aleatoria de un conjunto de datos (*bootstrapping*).
- Intercambiar marcadores de puntos de datos al realizar tests de significancia (prueba de aleatorización exacta).
- Validar modelos para el uso de subconjuntos aleatorios (*bootstrapping*, validación cruzada).

Entre las técnicas comunes de remuestreo [Chong,2003] se encuentran las:

- **Prueba de aleatorización exacta**. Es un procedimiento en el que los datos son reasignados aleatoriamente de manera que un valor exacto *p* se calcula basándose en los datos permutados.
- **Validación Cruzada**. En este proceso una muestra se divide aleatoriamente en dos o más subgrupos y resultados de pruebas se validan mediante la comparación de sub-muestras. El objetivo de este enfoque es saber si el resultado es replicable o sólo una cuestión de fluctuaciones aleatorias.
- ***Jackknife***. Es un paso más allá de la validación cruzada. En *Jacknife* se repite la misma prueba dejando un objeto, por lo que a esta prueba también se llama dejar un cabo.

- ***Bootstrap***. Significa que una muestra disponible da lugar a muchos otros ejemplos para el remuestreo. En este procedimiento, la muestra original podría ser duplicada tantas veces como los recursos de computación lo permitan, y a continuación, esta muestra expandida es tratada como una población virtual. Si bien el objetivo original de la validación cruzada es para verificar la replicabilidad de los resultados y la de *jackknife* es la detección de valores atípicos, *bootstrap* fue desarrollado con fines de inferencia.

2.6. Resumen del Capítulo

En este capítulo se presentaron algunas de las técnicas de clasificación plana, clasificación jerárquica y sobremuestreo que sirvieron como herramienta para el desarrollo de este trabajo de tesis. En este trabajo utilizamos la técnica de *Resampling* en conjunto con una técnica de creación de ejemplos artificiales, la cual se explicará con mayor detalle en los capítulos posteriores, para resolver el problema de desbalance de clases. Además, se describió el método de clasificación jerárquica multidimensional el cual se utilizó como parte del método propuesto.

Capítulo 3

Conceptos Astronómicos para la Clasificación de Galaxias

Desde sus inicios, la humanidad se ha dedicado a observar el cielo y a buscar una respuesta a los fenómenos que se presentan en éste, de ahí surge la astronomía como una ciencia que se dedica a estudiar los cuerpos celestes del universo, como son las estrellas, planetas y galaxias, entre otros. En sus inicios, la astronomía era un estudio sólo visual del Universo, ya que únicamente se contaba con los ojos del ser humano para percibir los fenómenos que ocurrían. Con el desarrollo de la tecnología, fue posible captar de manera más precisa una mayor cantidad de información, además de que se lograron observar objetos que se encuentran a grandes distancias y que el ojo humano no es capaz de percibir.

3.1. Objetos de Investigación en Astronomía

El Universo se compone de diversos objetos estelares, de los cuales la astronomía se encarga de estudiar; éstos pueden ir desde algo tan pequeño como una molécula hasta objetos inmensamente grandes como es el caso de las galaxias. Algunos de los principales objetos de interés para los astrónomos son:

La *Tierra* que es uno de los objetos de investigación para los astrónomos, los cuales se dedican a realizar observaciones de la atmósfera terrestre así como de los fenómenos

que ocurren fuera de ésta.

La *Luna*, a pesar de ser un objeto al cual el ser humano es capaz de llegar no deja de ser un objeto de investigación para los astrónomos, ya que es fácil de observar.

El *Sistema Solar*, en el cual se encuentra nuestro planeta. Éste pertenece a la galaxia espiral Vía Láctea y se encuentra en el brazo espiral conocido como Brazo de Orión. Está formado por una única estrella, el Sol y ocho planetas que orbitan alrededor de él.

El *Sol*, que es nuestra estrella más cercana y que además nos proporciona su energía, la cual es necesaria para la existencia de la humanidad. Por otro lado proporciona información para poder entender las condiciones de otras estrellas similares.

Las *Estrellas* pueden verse como puntos de luz, éstas se logran observar utilizando un telescopio y a su vez se clasifican, de acuerdo a las características que presentan, en gigantes, enanas, enanas blancas, etc.

Las *Galaxias*, que pueden verse como conjuntos masivos de estrellas, las cuales son el objeto de investigación para esta tesis.

3.2. Galaxias

Las galaxias son la base de la construcción del Universo. Una galaxia es una agrupación de miles de millones de estrellas, nubes de gas, planetas, polvo y materia oscura, unidos gravitacionalmente [Karttunen et al., 2007]. Éstas adoptan una gran variedad de formas y tamaños. Se estima que existen más de cien mil millones (10^{11}) de galaxias en el universo observable.

3.3. Esquemas de Clasificación de Galaxias

Existen dos formas de abordar el problema de la clasificación de galaxias: morfológica y espectral.

- El tipo morfológico se lleva a cabo a través de la descripción de la apariencia de la galaxia.
- El tipo espectral es una medida de la composición estelar, utilizando información de la luz dispersada en sus componentes básicos.

3.3.1. Clasificación Espectral de Galaxias

Actualmente, la mayoría de los grandes telescopios cuenta con espectrómetros, que son usados para medir la composición química y propiedades físicas de los objetos astronómicos, o para medir sus velocidades a partir del efecto Doppler de sus líneas espectrales. La función principal del análisis espectral es detectar la absorción o emisión de radiación electromagnética en ciertas longitudes de onda. La clasificación espectral de las galaxias proporciona información con respecto a la población estelar que la compone. Esto quiere decir que el espectro que integra a toda la galaxia se obtiene de la composición individual de las estrellas que la componen y de la absorción y emisión del medio interestelar. En la Figura 3.1 se puede observar un ejemplo de un espectro.

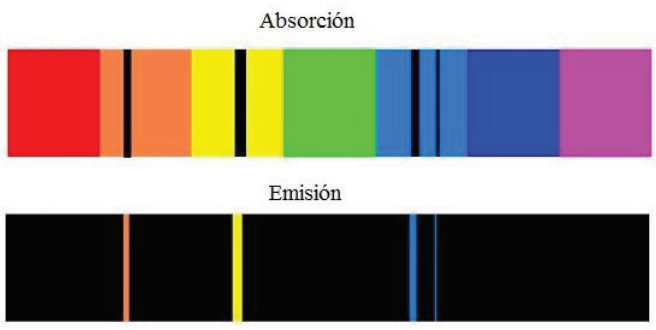

Figura 3.1: Espectro de absorción y emisión de un mismo elemento

Las galaxias elípticas y la mayoría de las lenticulares se consideran como *galaxias*

dominadas por líneas de absorción, mientras que las galaxias de tipo espiral e irregulares se consideran como *galaxias dominadas por líneas de emisión* como se menciona en el trabajo de [Coenda, 2008].

3.3.2. Clasificación Morfológica de Galaxias

Las galaxias tienen diversas morfologías, por lo que Edwin Hubble en 1926 ideó un método básico para la clasificación de ellas utilizando su forma [Karttunen et al., 2007]. Esta clasificación originalmente se consideraba como evolutiva, y así lo mostraba su diagrama. Las elípticas y S0 eran del tipo temprano, y las espirales e irregulares del tardío. Aunque ahora se considera errada esta idea, la terminología ha permanecido.

En su esquema de clasificación, hay tres tipos principales de galaxias: espirales, elípticas e irregulares, como se puede observar en la Figura 3.2.

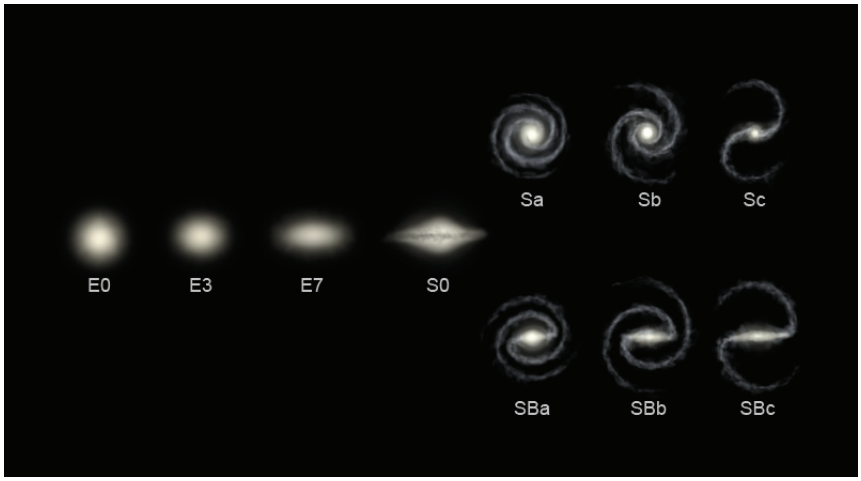

Figura 3.2: Diagrama diapasón de Hubble. Muestra los tres tipos principales de galaxias y sus respectivas subclases. Considerando de E0-E7 como galaxias elípticas, a Sa, Sb y Sc como galaxias espirales normales, a SBa, SBb y SBc como galaxias espirales barradas y finalmente a S0 como galaxias lenticulares.

Para la clasificación usaba tres criterios:

- La preponderancia relativa del bulbo en la imagen.
- El grado de apertura de los brazos espirales.

- El grado de resolución (estructura) de dichos brazos.

Las galaxias elípticas cuentan con una sub-clasificación que va de E0 a E7, según alargamiento de la elipse (en este caso E0 representa una galaxia circular, mientras que E7 representa una galaxia completamente elíptica). Las galaxias espirales se subdividen en ordinarias, barradas y lenticulares; mientras que las galaxias irregulares no presentan una forma elíptica o espiral que sea fácil de apreciar.

Una vez que se diera a conocer la secuencia propuesta por Hubble, diversos astrónomos llevaron a cabo el refinamiento de la misma. Tal es el caso de Gerard de Vaucouleurs como se menciona en [Eskridge et al. 1999]. Estos refinamientos se dieron principalmente en la clasificación de las galaxias espirales, introduciéndose los tipos intermedios E+ (un término medio entre elípticas y lenticulares), S0- (galaxias lenticulares las cuales sólo presentan la elipse mediante un estudio detallado), S00 (galaxias lenticulares con cierta estructura), S0+ (galaxias intermedias entre una lenticular y una Sa), Sab (entre Sa y Sb), Sbc (entre Sb y Sc), y Scd (entre Sc y Sd), así como la clasificación «*Pec*» (peculiar) para aquellas galaxias inclasificables (por ejemplo M82) y las galaxias enanas. El refinamiento más elaborado es el de Vaucouleurs ya que en éste se extendió la clase Sd incluyendo los tipos Sdm y Sm.

A continuación se llevará a cabo una descripción más detallada de las características morfológicas de los principales tipos de galaxias.

- *Galaxias Elípticas*. Éstas aparecen en forma de concentraciones elípticas de estrellas, las cuales son más brillantes en el centro y este brillo disminuye gradualmente hacia afuera. Las galaxias elípticas difieren entre ellas solamente en la forma, es decir en el alargamiento de la elipse, y sobre esta base se clasifican como E0, E1, ..., E7, cuanto más alto el número más elíptica, o sea, más alargada que ancha, como se puede observar en la Figura 3.3. Este número se puede conocer de la siguiente manera: Si los ejes menor y mayor de una galaxia elíptica son a y b respectivamente, entonces el tipo En esta definido como [Karttunen et al., 2007]:

$$n = 10\left(1 - \frac{b}{a}\right) \tag{3.1}$$

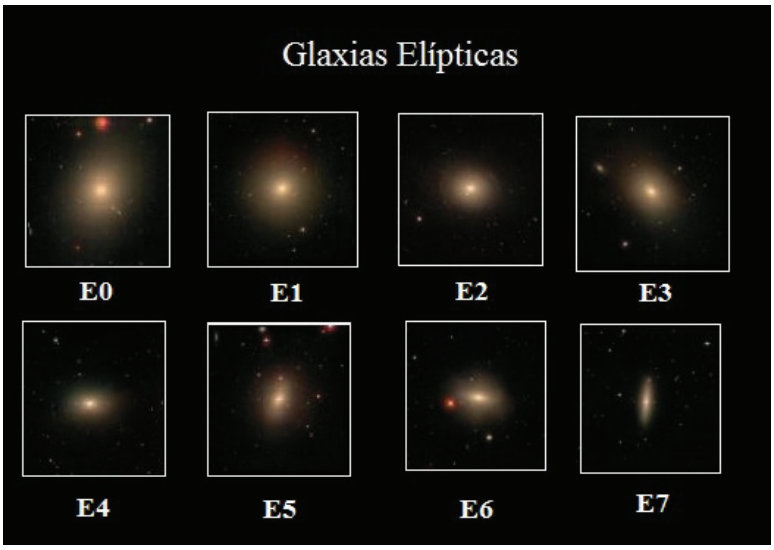

Figura 3.3: Ejemplo de los diferentes tipos de galaxias elípticas.

- *Galaxias Lenticulares*. Este tipo de galaxias se encuentran entre las elípticas y las espirales. Éstas no muestran una estructura en espiral pero contienen materia interestelar como en el caso de las elípticas. También contienen un disco plano formado por estrellas como las espirales. Además, muestran una componente difusa de disco (lente), como se puede observar en la Figura 3.4.

Figura 3.4: Galaxia lenticular.

- *Galaxias Espirales*. La característica principal de este tipo de galaxias es un patrón en espiral definido en el disco. Las galaxias espirales constan de un núcleo central que es similar al de las galaxias elípticas y un disco como en el caso de las lenticulares. Además de lo anterior, hay un disco que contiene gas y materia interestelar.

En este disco se forman nuevas estrellas, las cuales forman la espiral de la galaxia. Es decir, tienen un bulbo central del que arrancan los brazos espirales que están contenidos en un disco. Estas galaxias se dividen en dos grupos principales:

- **Ordinarias o Normales**. Estas galaxias están divididas en tres principales tipos dependiendo de qué tan apretados tengan sus brazos: Sa, Sb y Sc. Las galaxias Sa tienen los brazos muy apretados alrededor de un núcleo de la galaxia. Las galaxias Sc tienen los brazos muy sueltos alrededor del núcleo. Las galaxias Sb están en medio, teniendo los brazos moderadamente apretados alrededor del núcleo .
- **Barradas**. Éstas muestran una estructura lineal (barra) centrada en el núcleo, de donde salen los brazos espirales y que abarcan de un extremo a otro de la galaxia. A este tipo de galaxias se les conoce como **SB** y éstas a su vez se dividen en tres categorías principales SBa, SBb y SBc, las cuales difieren por la unión en sus brazos como en el caso de las ordinarias.

En la Figura 3.5 se pueden observar algunos ejemplos de los diferentes tipos de galaxias espirales que se mencionaron anteriormente.

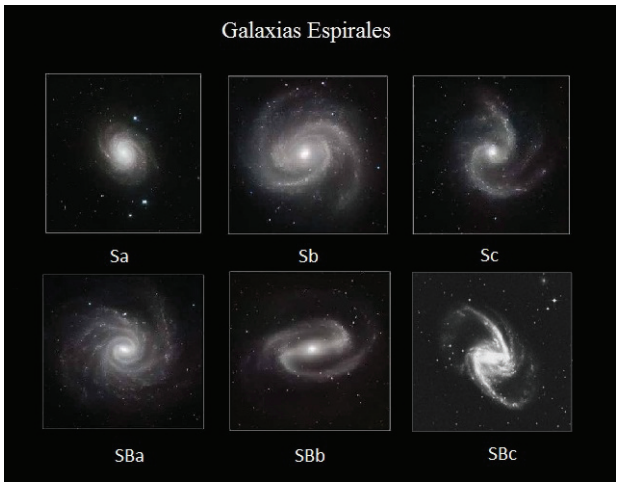

Figura 3.5: Ejemplo de los diferentes tipos de galaxias espirales.

- *Galaxias Irregulares*. Este tipo de galaxia no muestra ninguna simetría. Se dividen en dos tipos principales:

- Irr-I : En éstas hay presencia de un disco y una estructura espiral perturbada. Un ejemplo de este tipo de galaxia se puede observar en la Figura 3.6.

Figura 3.6: Galaxias irregular de tipo Irr-I.

- Irr-II: Presentan una estructura caótica, como se observa en la Figura 3.7.

Figura 3.7: Glaxias irregular de tipo Irr-II.

3.3.3. Características Relevantes de las Galaxias

Para llevar a cabo la clasificación morfológica de las galaxias, se han presentado diversos trabajos que proporcionan una lista de características que ayudan a discernir entre los diversos tipos de galaxias. Estas características se consideran con respecto a su morfología o estructura, por lo que se puede obtener de ellas el área, perímetro, la elipticidad, la asimetría, entre otras.

A continuación se definirán algunas de las características que se utilizaron en este trabajo de investigación para llevar a cabo el reconocimiento de galaxias en placas digitalizadas.

- *Área*. Es el número de pixeles que contiene el objeto. Considere que la función $I_n(i,j)$ describe el mapa de los objetos etiquetados de una imagen de tamaño MxN [Qiang Wu et al., 2002], donde i y j representan las coordenadas (x,y) dentro de la imagen.

$$I_n(i,j) = \begin{cases} 1 & \text{si } I(i,j) = n\text{-ésimo objeto,} \\ 0 & \text{en caso contrario} \end{cases}$$

  El área en pixeles para el n-ésimo objeto está dada por:

$$A_n = \sum_{i=1}^{M} \sum_{j=1}^{N} I_n(i,j) \tag{3.2}$$

  En términos de momentos geométricos ésta corresponde al momento de orden cero.

- *Elipticidad*. Es la Relación entre la longitud del semieje mayor y el semieje menor. En términos de momentos geométricos está dada por:

$$E = \frac{(\mu_{02} - \mu_{20})^2 + 4\mu_{11}}{Area} \tag{3.3}$$

  Donde μ_{02}, μ_{20} y μ_{11} representan los momentos de orden dos los cuales se explicaran en el siguiente capítulo.

- *Asimetría*. El término de asimetría fue introducido por [Abraham et al., 1994]. Ésta se define como la comparación entre la galaxia original y la galaxia rotada 180^o. La asimetría de una imagen está dada por la siguiente función:

$$A = \frac{\sum_{i,j} |I(i,j) - I_{180}(i,j)|}{\sum_{i,j} |I(i,j)|} - B_{180} \tag{3.4}$$

  Donde I_{180} es la imagen rotada y B_{180} es el promedio del fondo de la imagen rotado 180^o. Además i y j representan las coordenadas (x,y) dentro de la imagen.

- m_{q2q3}. Es la razón entre la pendiente ajustada de I(r) vs r para el segundo y tercer cuartil. Dividiendo el rango de r en cuatro segmentos iguales. Donde I(r) representa la intensidad de los pixeles en el radio r. La gráfica de la Figura 3.8 proporciona un ejemplo más claro.

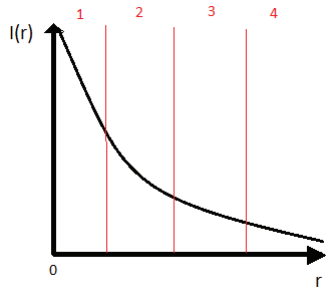

Figura 3.8: En la gráfica se representa la relación entre I(r) que es la intensidad de los pixeles con respecto de r que es el radio, se puede observar que se encuentra dividida en cuatro partes, estas divisiones representan los cuartiles. Para esta característica se consideran el segundo y tercer cuartil.

- r_{25}/r_{75}. Razón entre el radio que contiene el 25 % y el 75 % de la luz de la galaxia en la trama de I(r).

- R_{Buge}. Radio, donde I(r) cae el 90 % del valor máximo.

- *Concentración*. Mide el radio de la luz de la apertura interna dentro de una apertura externa, y está dada por la siguiente función:

$$C = 5\log(\frac{r_{80}}{r_{20}}) \tag{3.5}$$

  Donde r_{80} y r_{20} son las aperturas circulares que contienen el 80 % y el 20 % del flujo total de la galaxia, respectivamente.

- C_3,C_6. Los índices de concentración son una medida de la fracción de luminosidad total que cae dentro de cierto radio de elipticidad corregida. Éstos se definen de diversas formas en la literatura, pero en nuestro caso se consideran de la siguiente forma:

$$C_i = \frac{\int_0^{\alpha_i} I(r)dr}{\int_0^1 I(r)dr} \tag{3.6}$$

  Donde $\alpha_i = 0{,}1 - 0{,}9$

- *Entropía*. La entropía mide la suavidad de la imagen en términos de los valores de los niveles de gris. Entre más alto sea el nivel de entropía existen más niveles de gris en la imagen. La entropía puede ser calculada como:

$$Entropia = -\Sigma_{i,j} c_{ij} \log(c_{ij}) \tag{3.7}$$

- *Suavizado*. El parámetro de suavizado cuantifica el grado de la estructura de menor escala y está dado por la siguiente función.

$$S = \frac{\Sigma_{i,j} |I(i,j) - I_s(i,j)|}{\Sigma_{i,j} |I(i,j)|} - Bs \tag{3.8}$$

  Donde I_s es la galaxia suavizada y B_s es el promedio del fondo suavizado.

3.4. Trabajo Relacionado

La clasificación de galaxias es un problema que ha llevado a la realización de diversos trabajos de investigación, los cuales han abordado el problema de diversas maneras y utilizando diferentes técnicas tanto de clasificación como de extracción de características. En esta sección se hablará de algunos de ellos, de las técnicas que utilizaron y de los resultados que obtuvieron.

Para llevar a cabo la clasificación de galaxias, los primeros trabajos comenzaron a realizar la extracción de características que ayudarían a discernir entre los diferentes tipos de galaxias, un ejemplo de estos trabajos es el de [Bazell,2000]. En este trabajo se hace una investigación sobre la utilidad de un conjunto de características en el desempeño de clasificación de galaxias, utilizando *redes neuronales* con *retro-propagación*. Sus experimentos los basaron en una muestra de 805 imágenes de galaxias clasificadas por [Naim et al.,1995]. Ellos obtuvieron un total de 22 características por cada imagen, las cuales usaron para entrenar la red neuronal. Después de realizar sus experimentos, se puede observar que existen algunas características con mayor importancia para llevar a cabo la clasificación, como es el caso del parámetro m_{q2q3}. También fue el caso de los índices de concentración, pues existían algunos de ellos que permanecían constantes en ciertos tipos de galaxias.

Uno de los trabajos que se enfocó en la clasificación de galaxias, del mismo autor que el anterior, es el de [Bazell et al., 2001]. En este trabajo se lleva a cabo la comparación entre tres algoritmos de clasificación utilizando una muestra de 800 galaxias de un conjunto original de 834 imágenes tomadas de *Digitized Sky Survey*. Inicialmente, se extrajeron las 22 características de [Bazell,2000], pero la matriz de correlación usada mostró que un número de estas características están correlacionadas significativamente, por lo que se redujeron a solo 14 características ya que sólo se consideraron los índices de concentración tres y seis. Utilizaron un clasificador *Naive Bayes*, una red neuronal con retropropagación y un algoritmo de árboles de decisión con poda. Se considera un ensamble como un conjunto de clasificadores que de alguna manera combinan sus decisiones finales; normalmente esto lo hacen por medio de votación. Los ensambles a menudo dan mejores resultados que los clasificadores individuales. En este trabajo también se utilizaron los ensambles de redes neuronales, el clasificador *Naive Bayes* y J48. Para realizar estos ensambles se utilizó la técnica de *bootstrap aggregation*. En sus resultados se pueden observar que los ensambles obtuvieron mejores resultados que los clasificadores individuales. Los mejores resultados de clasificación se obtuvieron con el ensamble del algoritmo J48 ya que para todos los casos muestra los mejores porcentajes, por ejemplo para 2 tipos de galaxias presentó una clasificación correcta del 87.19 % pero al aumentar el número de tipos de galaxias este porcentaje descienden considerablemente, pues para 6 tipos de galaxias, la clasificación disminuye hasta el 45.95 %.

Otro trabajo que también se enfoca en la extracción de características de las galaxias es el trabajo de [Lotz et al., 2004]. Aquí se presentan dos métodos no paramétricos para la cuantificación morfológica de galaxias: la distribución relativa de los valores de los pixeles de la galaxia y el momento de segundo orden del $20\,\%$ de brillantez del flujo de la galaxia. El primero se conoce como el coeficiente **Gini** y el segundo método es conocido como M_{20}. El coeficiente Gini es una estadística basada en la curva de Lorenz, para la función de distribución acumulada de la riqueza de una población, o en este caso uno de los valores de los pixeles de una galaxia [Abraham et al.,2003]. M_{20} es definido como el momento de segundo orden del 20 % del flujo de la galaxia. Después se encargan de obtener tres características más de las galaxias las cuales son: la concentración, que la definen como la relación de los radios circulares que contiene 20 % y 80 % del

flujo total de la galaxia; la asimetría, que es la comparación entre la galaxia original y la galaxia rotada 180^o, y por último el suavizado que cuantifica el grado de la estructura de menor escala. Éstas fueron obtenidas de 170 imágenes en la frecuencia UV del óptico.

En el trabajo de [De la Calleja et al. 2004] se realizó un estudio experimental para así llevar a cabo la comparación de distintos algoritmos de clasificación enfocados a resolver el problema de la clasificación de galaxias. En este trabajo se lleva a cabo la clasificación de las galaxias por medio de su morfología, utilizando una base de datos de 292 imágenes de galaxias, la mayoría de ellas fueron tomadas del catálogo de la Sociedad Astronómica del Pacífico y su clasificación fue tomada del catálogo interactivo en línea NGC.

El método que ellos realizaron se divide en tres etapas: análisis de imágenes, compresión de los datos y por último la clasificación de galaxias. En la etapa de análisis de las imágenes, como bien se sabe, las imágenes con las que se trabaja varían tanto en tamaño, color y formato. Además, muchas veces la galaxia no se encuentra centrada dentro de la imagen, por lo que el objetivo en esta etapa fue crear imágenes invariantes a estos términos. Lo primero que hacen para alcanzar ese objetivo es encontrar las galaxias utilizando un umbral y así generar una imagen binaria. A continuación se obtiene la fila y la columna central de la galaxia, para después obtener la matriz de covarianza de los puntos de imagen de la galaxia. Por último, rotan la imagen para que el eje principal sea horizontal y así tener una mejor perspectiva de la galaxia que se desea analizar. Después se lleva a cabo la compresión de los datos utilizando el algoritmo *Principal Component Analysis* (PCA) que es un método estadístico que transforma una serie de (posiblemente) variables correlacionadas en un número (menor) de las variables no correlacionadas llamadas componentes principales (PC); es decir, permite reducir la dimensión del conjunto de datos, mientras retiene la mayor cantidad de información. En este caso utilizaron 8, 13 y 25 PC's ya que representa el 75 %, 80 % y 85 % de la información, respectivamente.

Una vez realizado el procedimiento anterior, se llevó a cabo la clasificación utilizando los algoritmos: *Naive Bayes, C4.5 y Random Forest*, y ensambles de cada uno de

ellos. Consideraron para un primer experimento con tres (E,S,Ir), después cinco (E, S0, Sa+Sb, Sc+Sd, Irr) y por último siete (E, S0, Sa, Sb, Sc, Sd, Irr) tipos de galaxias. De los tres clasificadores el que obtuvo mejores resultados para la clasificación de 3 tipos de galaxias fue *Random Forest* ya que llegó a un 91.64 % de clasificación correcta pero su precisión disminuyó considerablemente al incluir un mayor número de tipos de galaxias, pues se redujo al 43.62 % en el caso de siete galaxias.

Un trabajo más que trata la clasificación morfológica de galaxias es el trabajo de [De la Calleja et al.,2010]. En este trabajo se presentó un estudio experimental de seis algoritmos de aprendizaje automático aplicado a la clasificación morfológica de galaxias considerando el problema de desbalance de datos. Este problema se produce cuando el número de instancias de una clase supera por mucho las instancias de otra clase. En muchas ocasiones con conjuntos de datos desbalanceados los clasificadores obtienen mejor precisión en la clase mayoritaria pero baja precisión de predicción sobre la clase minoritaria que es generalmente de mayor de interés. Utilizaron los algoritmos de *Naive Bayes, Random Forest, Radial Basis Function Networks* (RBFNet), *C4.5, Support Vector Machines*. Para tratar el problema de desbalance utilizaron *The Synthetic Minority Over-sampling Technique* (SMOTE) y *Resampling*. Realizaron los experimentos con 310 imágenes para llevar a cabo la clasificación de tres tipos de galaxias y 293 imágenes para la clasificación de tres, cinco y siete tipos. De sus resultados se pudo concluir que *Random Forest* fue el mejor clasificador para la mayoría de los casos. Sin embargo, RBFNets y J48 obtuvieron buenos resultados. Además se puede mencionar que la técnica de *Resampling* obtuvo mejores resultados que *SMOTE* en casi todos los casos para todos los clasificadores. Las medidas que utilizaron para evaluar los métodos anteriores fueron las de *Recall*, *Precision* y *F-measure*. Para tres tipos de galaxias los mejores resultados se obtuvieron con el 100 % de sobre-muestreo, y los resultados fueron:*Recall*= 0.8825, *Precision*= 0.8919 y *F-measure*= 0.8615. Para cinco tipos de galaxias los mejores resultados se obtuvieron con el 500 % de sobre-muestreo, y los resultados fueron:*Recall*= 0.5236, *Precision*=0.5208 y *F-measure*= 0.5181. Por último para siete tipos de galaxias los mejores resultados se obtuvieron con el 500 % de sobre-muestreo, y los resultados fueron:*Recall*= 0.4231, *Precision*= 0.4632 y *F-measure*= 0.4335.

Todos los trabajos mencionados anteriormente se encuentran basados en un enfoque

de clasificación plana, utilizando técnicas como son ensambles de clasificadores y algunas de sobre-muestreo, pero se puede observar en todos los casos que al comenzar a aumentar el número de tipos de galaxias la precisión en la clasificación comienza a descender. Por lo que en este trabajo de tesis se probó un método de clasificación basado en un enfoque de clasificación jerárquica para obtener una mejor precisión a pesar de contar con varios tipos de galaxias.

3.5. Resumen del Capítulo

En este capítulo se presentaron algunos de los conceptos básicos de astronomía que son necesarios para llevar a cabo la clasificación de galaxias, como son: los tipos de galaxias y las características de cada uno de ellos, así como la diferencia entre clasificación espectral y clasificación morfologíca. Igualmente se presentó la clasificación de Hubble, la cual fue necesaria para diseñar la jerarquía que se utilizó en la fase de clasificación del método propuesto en este trabajo de tesis.

Además, se describieron algunos de los trabajos que se han enfocado al problema de la clasificación morfológica de galaxias, los cuales utilizan diferentes técnicas para resolver dicho problema, como son ensambles y algunas técnicas de sobre-muestreo, pero en todos los caso se muestra que al aumentar el número de tipos de galaxias la precisión en la clasificación desciende considerablemente.

Capítulo 4

Procesamiento de Imágenes

Las técnicas de procesamiento y análisis de imágenes son relativamente recientes. Estas técnicas tienen sus inicios hace unos 30 años aproximadamente, y han evolucionado muy rápidamente debido a las computadoras y su potencia de cálculo.

El procesamiento de imágenes consiste en aplicar técnicas automáticas de análisis de imagen mediante una computadora. En este capítulo veremos diferentes técnicas necesarias para llevar a cabo el procesamiento y análisis de las imágenes astronómicas.

4.1. Segmentación de Imágenes

Comúnmente nos referimos al método de segmentación de imágenes digitales como al proceso de dividir la imagen en diversas partes u objetos que la conforman. Esta división depende de la aplicación, pues se puede dividir la imagen en cuantas partes sea necesario para así detectar todos los objetos de interés.

Para llevar a cabo la segmentación de las imágenes se consideran dos conceptos básicos:

- *Similitud*. Los píxeles relacionados a un objeto deben ser similares respecto a algún criterio (color, nivel de gris, textura, etc.).
- *Discontinuidad*. Los objetos tienen formas geométricas que se definen por medio de contornos. Estos bordes distinguen unos objetos de otros.

Algunas de las técnicas más utilizadas dentro de la segmentación de imágenes son las siguientes:

- *Detección de bordes*. Es el procedimiento más empleado para la detección de discontinuidades. Un borde puede ser definido como la frontera entre dos regiones.

- *Umbralización*. Se emplea cuando hay una clara diferencia entre los objetos de interés con respecto al fondo de la imagen. Formalmente, una manera de separar los objetos del fondo consiste en seleccionar un umbral T que separe esos modos. Entonces, cualquier punto (x, y) para el que se cumpla que

$$I(x, y) > T \tag{4.1}$$

  se etiqueta como objeto; en otro caso, como fondo. Es decir, al aplicar un umbral, la imagen de niveles de grises quedará binarizada; etiquetando con '1' los píxeles correspondientes al objeto y con '0' aquellos que son del fondo.

- *Segmentación orientada a regiones*. Esta técnica de segmentación se basa en los criterios de similitud y continuidad de los píxeles que forman una región. Es decir, se considera que la imagen está formada por *n* regiones disjuntas, cada una de las cuales agrupa a los píxeles por alguna propiedad que los hace ser característicos de esa zona y distintos con respecto al resto.

4.2. Extracción de Características

Una vez que la imagen ha sido segmentada se debe de llevar a cabo una descripción de sus partes para poder identificar el objeto que se encuentra contenido en ella. El tipo de características que se extraen de los objetos depende de la aplicación. Éstas pueden ser con respecto a su morfología o bien a su estructura, por lo que puede obtenerse el área, perímetro, forma, color, etc. [Qiang Wu et al., 2002].

Existen diversas formas de llevar a cabo la extracción de características, una de estas formas es utilizando los momentos geométricos, los cuales se utilizaron en este trabajo

de investigación y se explicarán a continuación.

4.2.1. Momentos Geométricos

Los momentos geométricos son propiedades numéricas que se pueden obtener de una determinada imagen. Éstos nos proporcionan información de una imagen y tienen la ventaja de que no solo toman en cuenta los bordes sino que también cuentan los píxeles de la región de interés. Estos momentos nos sirven para reconocer una forma dentro de una imagen, como en el caso de la Figura 4.1

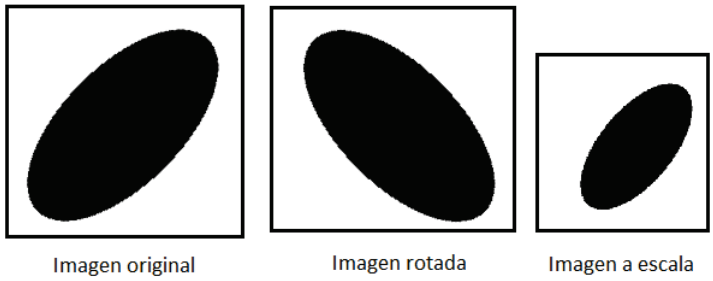

Figura 4.1: Los momentos geométricos son valores invariantes a rotación, traslación o escalamiento del objeto. En este caso se obtienen los valores de la elipse no importando si se le ha aplicado alguna transformación geométrica.

Existen diferentes momentos geométricos y de diferente orden, los cuales se detallan a continuación:

Momentos Simples

Si tenemos un objeto ubicado en una región de la imagen y éste está formado por los puntos *f(x,y)*, se define el momento *p,q* como:

$$m_{pq} = \int \int x^p y^q f(x,y) dx dy \tag{4.2}$$

para p,q=1,2,...

Ahora bien, se tiene una imagen definida por *f(x,y)*, donde *(x,y)* representan las coordenadas de la imagen y *f(x,y)* el valor que se encuentra en ese punto. En este caso será 0 si el punto es distinto de negro y 1 si es negro, dado que se trabaja con imágenes inversas. Si nuestra imagen no fuera binaria, entonces se tomarían en cuenta los valores que corresponden a cada uno de los puntos. Al ser imágenes digitales, el momento de orden *(p+q)* se define como [1]:

$$M(p,q) = \sum_x \sum_y x^p y^q f(x,y) \tag{4.3}$$

Momentos Centrales

Los momentos geométricos centrales se utilizan para identificar una figura dentro de una imagen, independientemente de su posición dentro de la imagen. Además, estos son invariantes a transformaciones geométricas de rotación, escalamiento y traslación. Para calcular estos momentos se utiliza el centroide *(X,Y)* de la figura. Los momentos centrales se definen de la siguiente manera:

$$MC(p,q) = \sum \sum \int (x - \overline{X})^p (y - \overline{Y})^q f(x,y) \tag{4.4}$$

Para calcular el centroide de la figura se utilizan los momentos de orden 0 y de orden 1, y se lleva a cabo de la siguiente manera.

$$\overline{X} = M(1,0)/M(0,0) \tag{4.5}$$

$$\overline{Y} = M(0,1)/M(0,0) \tag{4.6}$$

Momentos Centrales Normalizados

Con los momentos centrales Normalizados se pueden identificar figuras dentro de una imagen independientemente de su tamaño. Estos momentos están dados de la siguiente manera:

[1] **M** hace referencia a los momentos geométricos simples, **MC** a los momentos geométricos centrales, **U** a los momentos geométricos de orden dos. Por útilmo, μ se refiere a los momentos geométricos de orden tres y sus subindices al igual que en los otros momentos dependen de la combinación de p y q.

$$MCN(p,q) = MC(p,q)/MC^{\beta}(0,0) \tag{4.7}$$

Donde:

$$\beta = ((p+q)/2) + 1 \tag{4.8}$$

Momentos de Orden Cero

Los momentos de orden cero (M(0,0)) se utilizan para calcular el área de la figura, la cual se puede considerar como la suma de todos los píxeles que comprenden al objeto. Estos momentos se calculan con la siguiente ecuación:

$$M(0,0) = \sum_{x}\sum_{y} f(x,y) \tag{4.9}$$

Momentos de Orden Uno

La definición física de los momentos de orden uno nos dice que corresponden al punto de un objeto que tiene la misma cantidad de píxeles, que conforman al objeto, en cualquier dirección [Martín, 2002]. En el caso de imágenes se utiliza como punto de referencia del objeto. Esto quiere decir que los momentos de orden uno se utilizan principalmente para calcular el centroide de la figura contenida en la imagen. Estos momentos se definen de la siguiente manera:

$$M(1,0) = \sum_{x}\sum_{y} xf(x,y) \tag{4.10}$$

$$M(0,1) = \sum_{x}\sum_{y} yf(x,y) \tag{4.11}$$

Momentos Centrales de Orden Uno

$$U(1,0) = \sum_{x}\sum_{y}(x-\bar{x})^{1}(y-\bar{y})^{0}f(x,y) \tag{4.12}$$

$$= M(1,0) - \frac{M(1,0)}{M(0,0)}M(0,0) \tag{4.13}$$

$$U(0,1) = \sum_x \sum_y (x-\bar{x})^0 (y-\bar{y})^1 f(x,y) \tag{4.14}$$

$$= M(0,1) - \frac{M(0,1)}{M(0,0)} M(0,0) \tag{4.15}$$

Los Momentos Centrales de Orden Uno, por definición son cero, es decir al evaluar la expresiones 4.12 y 4.14 siempre se obtiene como resultado cero. Los momentos centrales normalizados, por definición al igual que los momentos centrales, son cero.

Momentos de Orden Dos

En los momentos de orden dos es donde se lleva a cabo el análisis de las imágenes a través del reconocimiento de formas. La densidad de la figura se multiplica por distancias al cuadrado desde el centro de masa o centroide (Inercia).

$$U(p,q) = \sum_x \sum_y (x-\bar{x})^p (y-\bar{y})^q f(x,y) \tag{4.16}$$

Momentos Centrales de Orden Dos

- MC(2,0): Su valor es proporcional al valor de la componente horizontal de una figura.
- MC(0,2): Su valor es proporcional al valor de la componente vertical de una figura.
- MC(1,1): Usa tanto la componente horizontal como la vertical. Su valor puede ser positivo o negativo dependiendo de donde se encuentre la componente vertical; si la componente vertical se encuentra en los cuadrantes 2º y 4º, entonces será negativo, si por el contrario está en los cuadrantes 1º y 3º entonces, será positivo. Por lo que en una figura simétrica respecto a los ejes, MC(1,1) será 0.

Momentos de Orden Tres

Éstos sirven para llevar a cabo el cálculo de los momentos invariantes. A partir de los

momentos centrales normalizados de orden 2 y 3 se obtienen los momentos invariantes, es decir, aquellos que se mantienen invariantes ante transformaciones geométricas. A continuación se presentan las ecuaciones para obtener los momentos de orden tres.

$$\mu_{00} = M(0,0) \tag{4.17}$$

$$\mu_{20} = M(2,0) - \overline{X}M(1,0) \tag{4.18}$$

$$\mu_{02} = M(0,2) - \overline{Y}M(0,1) \tag{4.19}$$

$$\mu_{11} = M(1,1) - \overline{Y}M(1,0) \tag{4.20}$$

$$\mu_{30} = M(3,0) - 3\overline{X}M(2,0) + 2\overline{X}M(1,0) \tag{4.21}$$

$$\mu_{12} = M(1,2) - 2\overline{Y}M(1,1) - \overline{X}M(0,2) + 2\overline{Y}^2M(1,0) \tag{4.22}$$

$$\mu_{12} = M(2,1) - 2\overline{X}M(1,1) - \overline{Y}M(0,2) + 2\overline{X}^2M(0,1) \tag{4.23}$$

$$\mu_{03} = M(0,3) - 3\overline{X}M(0,2) + 2\overline{Y}^2M(0,1) \tag{4.24}$$

4.3. Resumen del Capítulo

En este capítulo se describieron algunas de las técnicas que son utilizadas para llevar a cabo la segmentación y la extracción de características de imágenes. El algoritmo de segmentación que se utilizó para el desarrollo de este proyecto de tesis es el de umbralización. Este algoritmo en general realiza una buena segmentación de estrellas y galaxias

en la placa astronómica [Stockman et al.,2001]. En el proceso de extracción de características se utilizaron los momentos geométricos para extraer información relevante de las imágenes para utilizarla en la clasificación de galaxias.

En el siguiente capítulo se describe la estrategia que se utilizó para resolver el problema de clasificación de galaxias en placas digitalizadas.

Capítulo 5

Clasificación Jerárquica de Galaxias

En este capítulo se describe la estrategia que se llevó a cabo para resolver el problema de clasificación de galaxias en placas digitalizadas basada en un enfoque jerárquico. El método propuesto está compuesto por 7 etapas, que inicia con el procesamiento de las placas hasta la validación del algoritmo de clasificación, como se muestra en la Figura 5.1.

5.1. Segmentación de Placas Astronómicas

El primer paso para llevar a cabo la clasificación de galaxias es la segmentación de objetos estelares, en particular estrellas y galaxias. Se analizaron diferentes técnicas de segmentación como lo son detección de bordes y umbralización. Como se mencionó en el capítulo anterior, el método utilizado para realizar la segmentación de las placas, es el de umbralización. Los métodos de umbralización son de los más eficientes y sencillos para segmentar así como para encontrar con mayor facilidad los objetos de interés dentro de la placa digitalizada [Stockman et al.,2001], tal y como se muestra en la Figura 5.3. Los métodos de umbralización utilizan el histograma para indicar el número de puntos en los que la imagen posee un determinado nivel de gris.

La umbralización trata de determinar un valor de intensidad, llamado umbral (*threshold*), que separa las clases deseadas; es decir, a partir de histogramas se elige nivel de gris que separa los valores correspondientes al objeto y al fondo. Su principal limitación es que no es fácil encontrar el umbral idóneo que permita encontrar todos los objetos

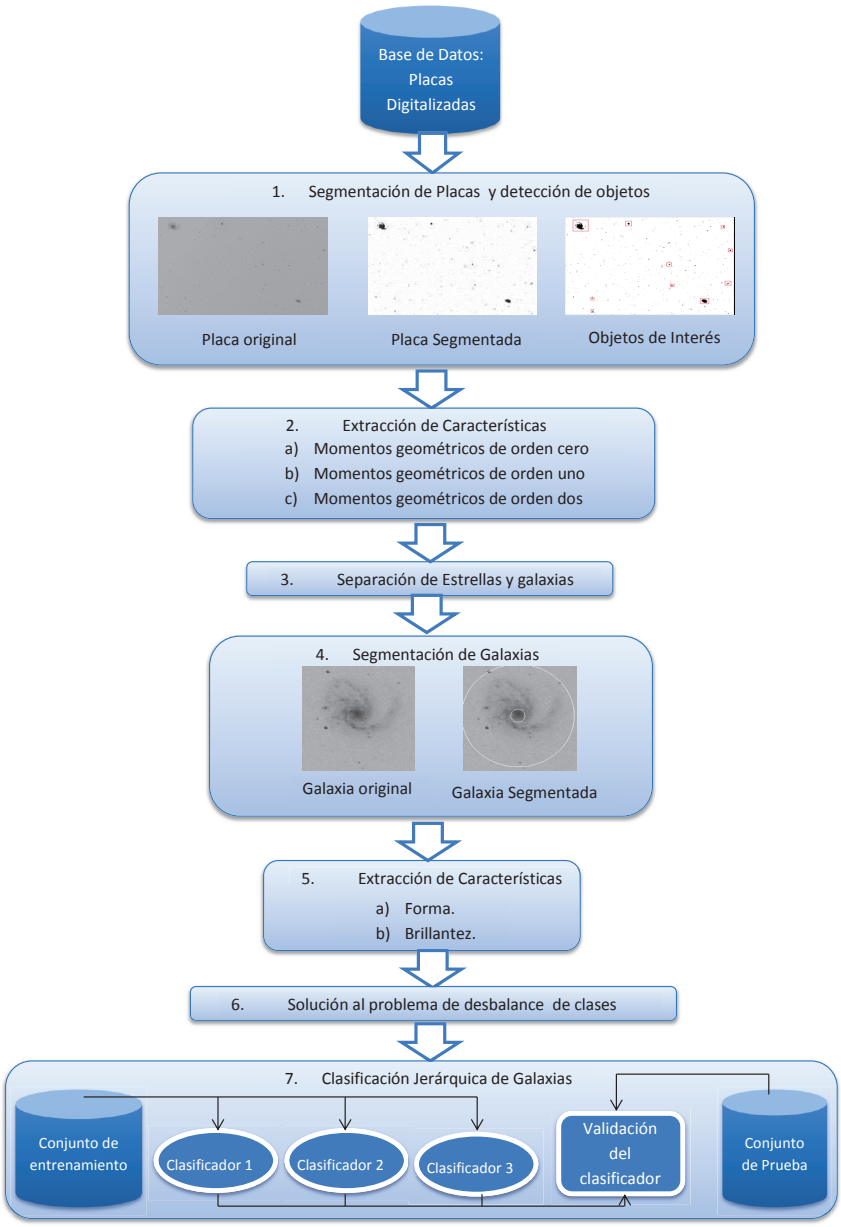

Figura 5.1: Esquema general del método de clasificación de galaxias a partir de placas digitalizadas. Una vez que se tiene la base de datos de placas digitalizadas, en las etapas 1 y 2 se realiza el procesaminto de las imágenes de las placas para así en la etapa 3 separar las estrellas de la galaxias. En los pasos 4, 5 y 6 se lleva a cabo el procesamiento de las imágenes de las galaxias, para finalmente en etapa 7 realizar la clasificación de los diferentes tipos de galaxias.

como se puede observar en la Figura 5.2, por lo anterior se utilizó un rango en el umbral de 120-150 para cada una de las imágenes de las placas y así poder acercarnos más al umbral deseado. Estos valores se obtuvieron de manera experimental.

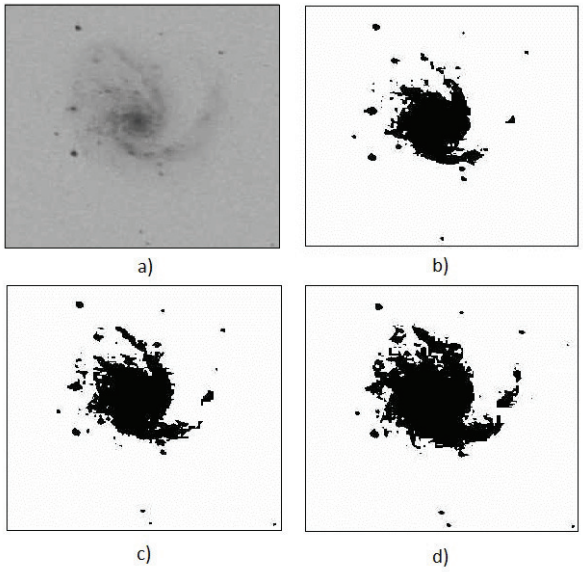

Figura 5.2: Ejemplo de segmentación por umbralización. a)Muestra la imagen original, b)Imagen segmentada con un umbral de 120, c)Imagen segmentada con un umbral de 130 y d)Imagen segmentada con un umbral de 150

Una vez que han sido detectados los objetos, se comenzó a trabajar con cada uno de ellos de manera independiente, y descartándose algunos de ellos bajo el criterio de área, eliminando aquellos objetos que contaban con una cantidad menor de 500 pixeles, por lo que solo nos enfocamos en la región de la imagen donde se encontraban los objetos de interés, ver Figura 5.3. Una vez realizado lo anterior se procedió a llevar a cabo la clasificación de estrellas y galaxias, procedimiento que se explica en la siguiente sección.

5.2. Clasificación de Galaxias/Estrellas

Una placa astronómica puede contener cientos de objetos estelares, entre los cuales comúnmente se presentan grandes poblaciones de estrellas. Por lo anterior, es necesario

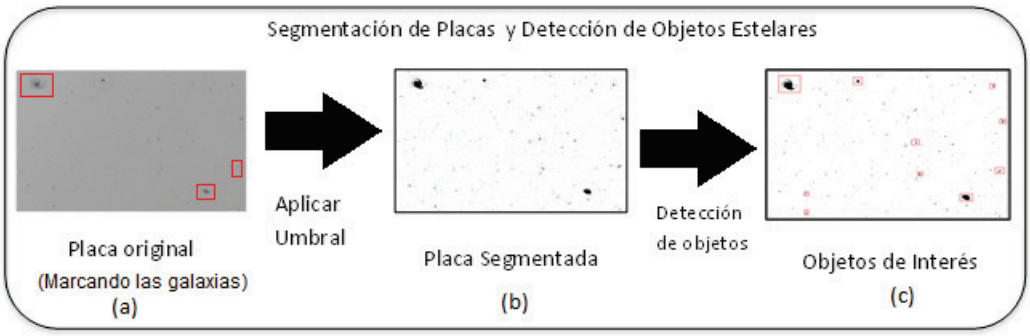

Figura 5.3: Segmentación de placas. (a) es el ejemplo de la imagen de una placa en donde podemos observar que ésta contiene tres galaxias, (b) es la placa segmentada utilizando una técnica de umbralización y por último (c) muestra los objetos de interés que se detectaron considerando solo el área como criterio de eliminación.

poder llevar a cabo la detección de las galaxias y así distinguir entre galaxias y estrellas dentro de la imagen. Así que lo que se hizo después de segmentar la imagen de la placa astronómica fue obtener los momentos geométricos de orden cero, uno y dos; de los objetos. Éstos nos ayudan a conocer características como el área, y la elipticidad del objeto, las cuales son utilizadas como atributos para llevar a cabo la clasificación. En esta etapa partimos de la idea de que las estrellas presentan una forma circular mejor definida que en el caso de las galaxias, dado que las estrellas son objetos puntuales y las galaxias son objetos extendidos. Se utilizaron los clasificadores *Naive Bayes* y *Random Forest* para llevar a cabo dicha clasificación, los cuales se encuentran implementados en la herramienta Weka [Witten et al.,2005].

5.3. Procesamiento de las Imágenes de Galaxias

Una vez separadas las estrellas de las galaxias, nos enfocamos únicamente en los objetos de nuestro interés, en este caso son las galaxias.

En las etapas anteriores ya se cuenta con la imagen segmentada, entonces lo que se hace es separar el núcleo del resto de la galaxia. De acuerdo con el trabajo de [Lotz et al., 2004] sabemos que en el núcleo de las galaxias se encuentra el 20 % del flujo de la misma. Por lo anterior es que se utilizaron los momentos geométricos de or-

den uno y cero para poder ubicarnos en el centro de la galaxia, ya que la razón de los momentos de primer orden entre los de orden cero proporciona el centroide de la figura. Una vez en el centro de la galaxia se calculó el radio de Petrosian [Petrosian,1982] para poder considerar solo el 20 % de la galaxia tal como se muestra en la Figura 5.4, lo cual se considera el núcleo de la galaxia y el 80 % restante se considera como el cuerpo resante de la misma.

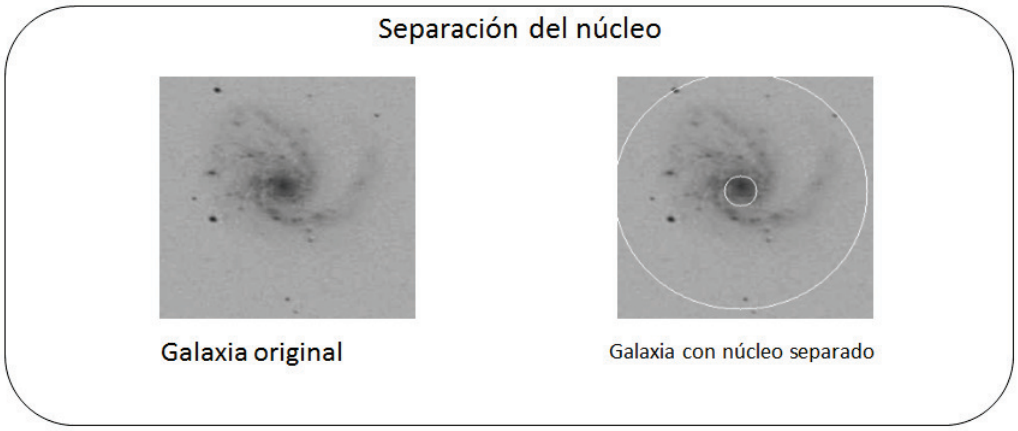

Figura 5.4: Separación del núcleo de la galaxia.

Lo anterior nos permite obtener características que se relacionan con la brillantez de la galaxia, como lo es $r_{25}r_{75}$, la razón que existe entre el brillo del núcleo de la galaxia y el resto de la misma.

5.3.1. Extracción de características

Una vez segmentada la imagen de cada galaxia, el siguiente paso que se realizó fue extraer las características descriptivas de la galaxia, las cuales se representan a través de un vector de atributos $(x_1,x_2, x_3, ..., x_n)$. Estas características se detallaron en los capítulos 2 y 3 y se listan en la tabla 5.1.

Características descriptivas extraídas de las imágenes de galaxias	
M(0,0)	Momentos geométricos de orden cero.
M(0,1),M(1,0)	Momentos geométricos de orden uno.
M(2,0),M(0,2), M(1,1)	Momentos geométricos de orden dos.
Máximo brillo	Nivel de brillo máximo en la imagen.
$\mathbf{m_{q2q3}}$	Relación entre la pendiente ajustada de I(r) vs r para el segundo y tercer cuartil.
Elipticidad	Relación entre la longitud del semieje mayor y el semieje menor.
Área	Número de pixeles contenidos en el objeto.
Max(rL)	El valor máximo de la trama de rI(r) vs r.
Asimetría	Comparación entre la galaxia original y la galaxia rotada 180^0.
$\mathbf{r_{25}r_{75}}$	Razón entre el radio que contiene el 25 % y 75 % de la luz de la galaxia.
$\mathbf{R_{Bulqe}}$	Radio, donde I (r) cae al 90 % del valor máximo.
$\mathbf{c_3}, \mathbf{c_6}$	Índice de concentración en los anillos 3 y 6.
$\mathbf{P_{max}}$	Valor máximo de la normalización de co-ocurrencia de la matriz c_{ij}.
Entropía	$-\sum_{ij} c_{ij} log(c_{ij})$.
Concentración	Mide el radio de la luz de la apertura interna dentro de una apertura externa.
Suavizado	Cuantifica el grado de la estructura de menor escala.

Tabla 5.1: Características que se consideraron para realizar la clasificación de galaxias.

5.4. Generación de Ejemplos Artificiales

Existe un importante desbalance en cuanto al número de ejemplos que podemos encontrar de un tipo de galaxia a otro. Por ejemplo, existe un mayor número de galaxias espirales que de cualquier otro tipo de galaxias en el universo. Igualmente, existe un número muy reducido de galaxias irregulares. Por lo anterior, en esta etapa del método se crearon ejemplos artificiales para ayudar a disminuir ese problema de desbalance.

Los ejemplos artificiales se crearon a partir de las imágenes originales, a las cuales se les aplicaron transformaciones geométricas como son:

- **Rotación**. Que son cambios en la orientación. Ésta puede considerarse de dos formas. La primera es con respecto al origen, es decir, la posición de un punto es rotada alrededor del origen de coordenadas. La segunda forma es teniendo como referencia cualquier punto (x_c, y_c) en el plano. En general la función de rotación está dada como:

$$\begin{cases} x^{'} = x_c + (x - x_c)cos\theta - (y - y_c)sin\theta \\ y^{'} = y_c + (x - x_c)sin\theta) + (y - y_c)cos\theta \end{cases} \tag{5.1}$$

  La función de rotación realiza cambios en características como son los momentos geométricos de segundo orden ya que estos obtienen información sobre la orientación del objeto dentro de la imagen.

- **Escalado**. Esta función se refiere a cambios en el tamaño. Al igual que la función de rotación se obtiene de dos formas. La primera es con respecto al origen, en este caso la posición del punto se multiplica por una constante y hay que especificar los dos factores de escala S_x y S_y. En la segunda forma se debe considerar que si el origen de coordendas no se encuentra en el interior del objeto, entonces se produce un desplazamiento. Para evitarlo, se usa un punto fijo, este punto puede ser el centro del objeto o uno de sus vértices o bien un punto arbitrario. Por lo tanto la función general del escalado de un objeto estaría dada por la siguente ecuación:

$$\begin{cases} x^{'} = x_c + S_x(x - x_c) \\ y^{'} = y_c + S_y(y - y_c) \end{cases} \tag{5.2}$$

La función de escalado principalmente afecta a las características que tienen relación con el área del objeto, como son los momentos geométricos de orden cero.

Para la creación de los ejemplos artificiales se tomaron los ángulos de $45^o, 90^o$ y 180^o y el escalado se realizó en un rango del 50 % y 30 % dependiendo del tamaño de las galaxias, dado que experimentalmente se observó que estos parametros permiten conservar la mayor parte de la información de la galaxia. Un ejemplo de esto se puede observar en la Figura 5.5.

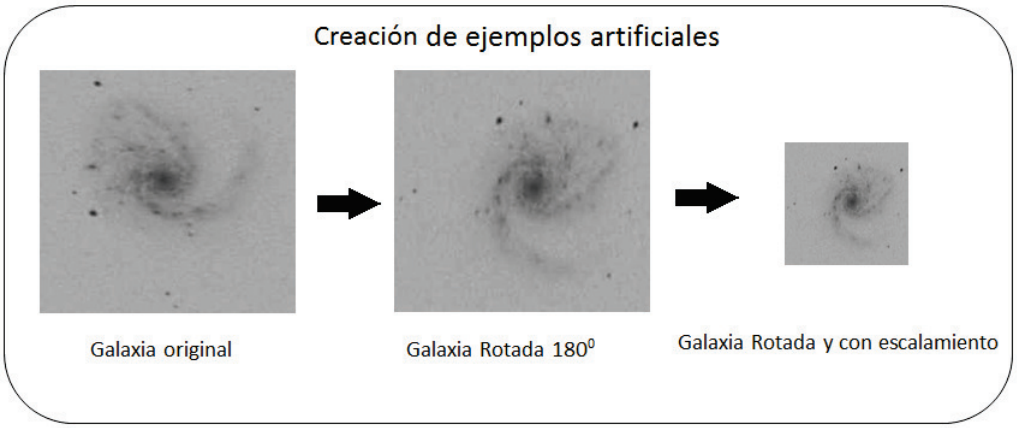

Figura 5.5: Creación de ejemplos artificiales a través de funciones geométricas.

También se utilizó el método de *resampling* [Chong,2003], el cual se encuentra implementado en el software **Weka**, y permite hacer sobre-muestreo de los ejemplos en el porcentaje que el usuario lo desee.

5.5. Clasificación Jerárquica de Galaxias

Para realizar la clasificación el primer paso fue determinar la jerarquía que se utilizaría para llevar a cabo nuestros experimentos, que como se había mencionado con anterioridad, se encuentra basada en el diagrama de diapasón de Hubble. Esta jerarquía

se puede observar en la Figura 5.6.

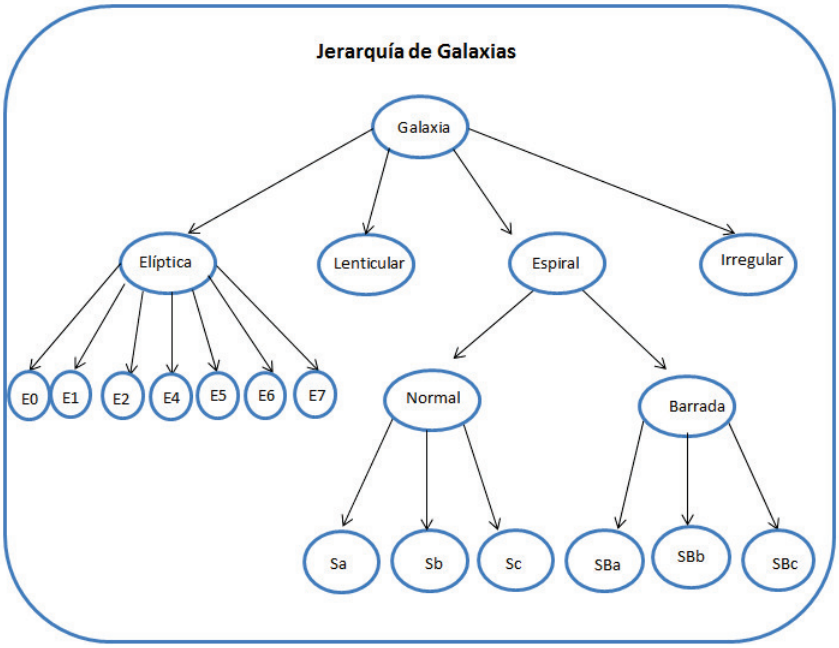

Figura 5.6: Jerarquía basada en el diagrama de Hubble para la clasificación de galaxias.

Dentro de los métodos de clasificación jerárquica uno de los que ha presentado mejores resultados es el método de clasificación jerarquica multidimensional [Hernandez,2012] por lo que se determinó que la mejor forma para abordar este problema de clasificación de galaxias era utilizando dicho método de clasificación. En este método de clasificación jerárquica se busca la forma de considerar la relación que existe entre los nodos, para llevar a cabo la clasificación. Dicha relación se basa en la combinación de las probabilidades predichas por cada clasificador. Como se mencionó anteriormente existen tres formas de llevar a cabo dicha combinación de probabilidades. En este trabajo de tesis se utilizó la técnica denominada producto de probabilidades (**PP**).

5.6. Resumen del Capítulo

En este capítulo se describió la estrategia que se propuesto para realizar la clasificación morfológica de galaxias. La estrategia de clasificación que se propuso involucra tres procesos, la segmentación de imágenes, la extracción de características y la clasi-

ficación. En el proceso de segmentación de imágenes se utilizó el algoritmo de umbralización y la separación del núcleo del resto del cuerpo de la galaxias. En la extracción de características se consideraron atributos presentados en el trabajo de [Bazell et al., 2001] y algunos que se obtuvieron utilizando los momentos geométricos.

Para la clasificación de galaxias se propone un clasificador jerárquico multidimensional, utilizando la variante de producto de probabilidades, en el cual se considera la relación que existe entre los nodos al multiplicar las probabilidades de las trayectorias lo cual ayuda a obtener una mejor precisión en la clasificación.

Capítulo 6

Experimentos y Resultados

En este capítulo se presentan los resultados de la evaluación experimental del método para la clasificación de galaxias utilizando el modelo jerárquico que se describió en el capítulo anterior. Se incluye una comparación entre los resultados obtenidos con la clasificación jerárquica y con una clasificación plana, además de los resultados de clasificación plana de cuatro y nueve tipos de galaxias.

También se presentarán los resultados que se obtuvieron al realizar la clasificación entre estrellas y galaxias. Finalmente se hace un análisis acerca de los resultados que se obtuvieron.

6.1. Creación de la Base de Datos

Para realizar los experimentos de este trabajo de tesis fue necesaria la creación de la base de datos. Esta base de datos está conformada por placas fotográficas astronómicas digitalizadas que fueron tomadas principalmente en el rango del azul del *visible* del espectro electromagnético, las cuales pertenecen a la colección de la cámara Schmitd del INAOE.

Debido a que estas placas aún no se encontraban digitalizadas, primero se utilizó el catálogo del observatorio [Tecpanecatl,2002] para poder identificar las galaxias que se encuentran contenidas en el acervo del instituto y así comenzar con la digitalización.

El siguiente paso fue la identificación de cada una de las galaxias dentro de cada placa para poder definir su tipo. El método se basa en una clasificación supervisada; es decir, una vez que la placa es digitalizada, utilizamos los mapas de Palomar para poder ubicar su posición; es decir, sus cordenadas (x,y) dentro de la placa. Una vez que la galaxia es localizada, ésta se busca en bases de datos astronómicas para poder saber a que tipo corresponde. Las bases astronómicas que utilizamos son la *Astronomical Database-Université de Strasbourg* (SIMBAD)y la *NASA/IPAC Extragalactic Database* (NED). Al final se obtuvo como resultado una base de datos de 132 imágenes de galaxias, las cuales se encuentran distribuidas como se muestra en la Figura 6.1.

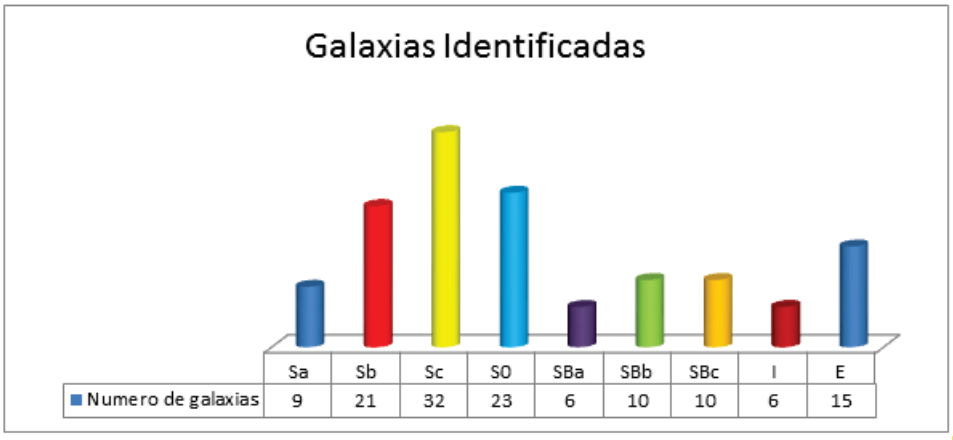

Figura 6.1: Distribución de Galaxias. La gráfica muestra el número de galaxias que se encontraron por cada tipo, contenidas dentro de las placas digitalizadas.

Desde este momento se puede observar que existe un gran desbalance entre las clases, por lo que se tuvo que buscar una técnica para poder resolver este problema. Esta técnica fue la creación de ejemplos artificiales utilizando transformaciones geométricas, la cual se describió en el capítulo anterior. En total se crearon 55 nuevos ejemplos los cuales se encuentran distribuidos como se muestra en la tabla 6.1. Estos ejemplos también se agregaron a nuestra base de datos.

Clase	Sa	SBa	SBb	I	E
Número de ejemplos Artificiales	11	14	10	15	5

Tabla 6.1: Ejemplos Artificiales

6.2. Clasificación de Galaxias y Estrellas

Como se mencionó en el capítulo anterior, dado que una placa puede contener cientos de objetos estelares, fue necesario hacer una primera clasificación para así poder discernir entre estrellas y galaxias. Esta clasificación se realizó sobre un total de 93 imágenes de objetos estelares. A estas imágenes se les calcularon los momentos geométricos sobre las imágenes binarizadas. Los momentos geométricos de orden cero, uno y dos fueron utilizados como atributos para llevar a cabo la clasificación.

Para llevar a cabo esta primera clasificación se utilizaron los clasificadores *Naive Bayes* y *Random Forest*, los cuales se encuentran implementados en el sofware de weka [Witten et al.,2005]. La precisión de la clasificación se obtuvo utilizando *10-fold cross validation* para cada método. Después de esta primera clasificación se obtuvieron los resultados que se muestran en la tabla 6.2.

Clasificador	**Porcentaje de precisión**
Naive Bayes	79.16
Random Forest	**91.66**

Tabla 6.2: Porcentaje de clasificación de estrellas y galaxias

En esta primera etapa se obtuvieron buenos resultados pues se logró hasta un 91.66 % de precisión en la clasificación entre estrellas y galaxias, utilizando el clasificador *Random Forest*.

6.3. Clasificación de Galaxias

En esta sección se presentarán los resultados obtenidos al realizar tanto la clasificación jerárquica de las galaxias así como la clasificación de tipo plana de las mismas.

6.3.1. Clasificación Plana de Galaxias

Para llevar a cabo la evaluación de todos los experimentos se utilizó la técnica *10-fold-cross validation* definida en el capítulo dos. Los resultados que se muestran a continuación fueron obtenidos al promediar los resultados de realizar cinco veces la validación cruzada. Estos resultados consideran la clasificación sin ejemplos artificiales, con ejemplos artificiales, así como utilizando *resampling* al 100 %, 300 % y 500 %.

En la tabla 6.3 se muestra el número de clases, así como los tipos de galaxias que se consideraron para realizar la clasificación en esta etapa.

Clases	**Tipos de galaxias**
4	E, S, S0, I
9	E, Sa, Sb, Sc, SBa, SBb, SBc, S0, I

Tabla 6.3: Tipos de galaxias considerados en la clasificación plana.

Los resultados de la clasificación plana para cuatro tipos de galaxias se pueden observar en la tabla 6.4.

En la tabla 6.5, se muestran los resultados de clasificación plana para nueve tipos de galaxias.

En la clasificación plana se llevaron a cabo 4 pruebas. En la primera prueba de la clasificación plana se utilizaron solo las imágenes que pertenecen a la base de datos que se creó anteriormente. Para la segunda prueba de la clasificación plana se utilizaron las

	Naive Bayes (%)	**Random Forest** (%)
Datos	25.5814	58.1395
Datos y Ejemplos Artificiales	33.65	58.65
Datos y Resampling (100 %)	40.45	60.55
Datos y Resampling (300 %)	36.38	58.19
Datos y Resampling (500 %)	38.88	64.03
Datos + Ejemplos Artificiales + Resampling (100 %)	34.99	57.99
Datos + Ejemplos Artificiales + Resampling (300 %)	34.77	61.87
Datos + Ejemplos Artificiales + Resampling (500 %)	33.60	**64.17**

Tabla 6.4: Porcentajes de precisión en la clasificación plana para cuatro tipos de galaxias.

	Naive Bayes (%)	**Random Forest** (%)
Datos	22.09	23.25
Datos y Ejemplos Artificiales	26.92	30.76
Datos y Resampling (100 %)	29.02	27.91
Datos y Resampling (300 %)	27.63	31.25
Datos y Resampling (500 %)	26.52	29.99
Datos + Ejemplos Artificiales + Resampling (100 %)	25.04	38.42
Datos + Ejemplos Artificiales + Resampling (300 %)	31.35	33.16
Datos + Ejemplos Artificiales + Resampling (500 %)	34.21	**39.44**

Tabla 6.5: Porcentajes de precisión en la clasificación plana para nueve tipos de galaxias.

imágenes que pertenecen a nuestra base de datos más los ejemplos artificiales. La tercera prueba se dividió en tres, ya que se trabajó con las imágenes de la base de datos y resampling de las mismas al 100 %, 300 % y 500 %. En la última prueba se utilizaron las imágenes de la base de datos junto con los ejemplos artificiales y además *Resampling*. En todas las pruebas se obtuvieron mejores resultados con el clasificador *Random Forest*. Además, el mejor de los porcentajes de clasificación se obtuvo al combinar las diferentes técnicas que se emplearon para la solución del desbalance de clases y utilizando el clasificador *Random Forest*, ya que se obtuvo hasta un 64.17 % de precisión con cuatro tipos de galaxias y 39.44 % con nueve tipos.

6.3.2. Clasificación Jerárquica de Galaxias

Como se ha mencionado con anterioridad, el algoritmo que se utilizó para llevar a cabo la clasificación jerárquica de galaxias, está basado en el trabajo de [Hernandez,2012]. Para esta clasificación se utilizó la jerarquía de la Figura 6.2.

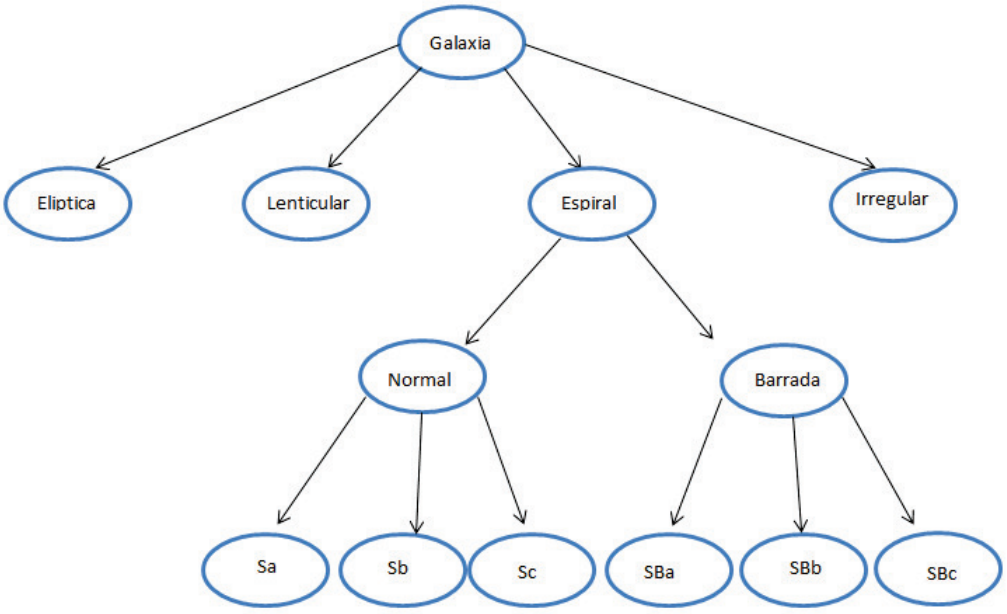

Figura 6.2: Taxonomía de galaxias utilizada para la clasificación jerárquica de galaxias.

Los resultados que se obtuvieron al realizar la clasificación jerárquica se observan en la tabla 6.6. Para llevar a cabo esta clasificación se separaron el conjunto de entrenamiento y el de prueba. Al conjunto de entrenamiento fue al único al que se le aplicaron las técnicas de sobre-muestreo. Para realizar la evaluación de estos experimentos, al igual que en la clasificación plana, se utilizó la técnica *10-fold-cross validation*.

	Naive Bayes (%)	**Random Forest** (%)
Datos	28.57	42.85
Datos + Artificiales	21.42	28
Datos + Resampling (500 %)	25	42.86
Datos + Artificiales + Resampling(500 %)	21.43	**53.57**

Tabla 6.6: Porcentajes de precisión en la clasificación jerárquica considerando nueve tipos de galaxias

Dado que en la clasificación plana los mejores resultados para cuatro y nueve tipos de galaxias se obtuvieron al utilizar tanto *resampling* al 500 % y de ejemplos artificiales, así como la combinación de ambos, se decidió que las pruebas para la clasificación jerárquica se realizaran sólo sobre esos conjuntos de datos.

Para la clasifican jerárquica también se realizaron cuatro pruebas. Para la primera prueba solo se consideraron las imágenes de nuestra base de datos y los clasificadores *Naive Bayes* y *Random Forest*. Para la segunda prueba se utilizaron tanto las imágenes ya mencionadas, como los ejemplos artificiales de las mismas. En el caso de la tercera prueba solo se utilizaron las imágenes a las que se les aplicó resampling al 500 %. Por último, se realizó una prueba con las imágenes, los ejemplos artificiales y con resampling al 500 %. En todos los casos el algoritmo de clasificación que presentó mejores resultados fue *Random Forest*, pues obtuvo hasta el 53.57 % de clasificación correcta para un total de nueve tipos de galaxias. Esta es una mejora significativa respecto a la clasificación plana que fue de 39.44 %, es decir más del 14 %.

6.4. Comparación con otros métodos

Después de llevar a cabo los experimentos utilizando la base de datos que se creo en este trabajo de tesis se llevaron a cabo los experimentos utilizando la base de datos descrita en [Bazell,2000]. A estas imágenes se le extrajeron las características descritas en la sección 5.3.1 y se les aplicaron las técnicas de sobre-muestreo (*Resampling* y la creación de ejemplos artificiales). En este caso solo se consideraron a seis tipos de galaxias (E, S0, Sa, Sb, Sc y Sd/Sm/Irr), dado que en el trabajo con el cual deseamos compararnos solo consideran estos tipos de galaxias. Por lo tanto la jerarquía que se utilizó se muestra en la Figura 6.3 y los resultados se muestran en la Tabla 6.7.

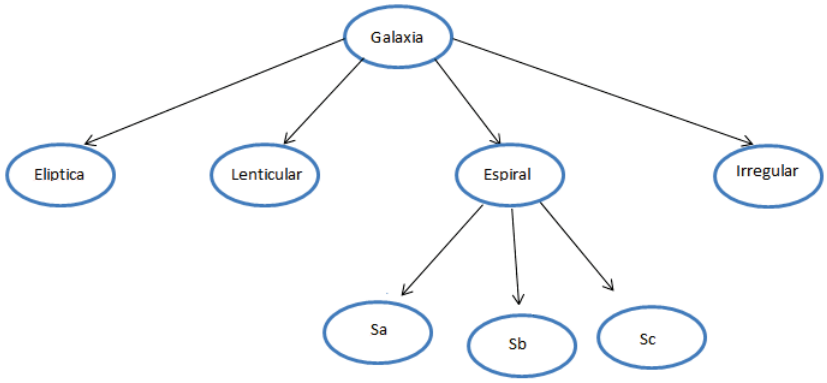

Figura 6.3: Jerarquía utilizada para los experimentos realizados con la base de datos de Bazell y Aha.

	6 clases
Bazell and Aha	45.95
Método propuesto	46.42

Tabla 6.7: Comparación del método jerárquico con otros trabajos

Como se puede observar en la Figura 6.3 la jerarquía solo cuenta con dos niveles, por lo tanto no existe gran diferencia entre la clasificación plana y la jerárquica y es por

ello que en los resultados que se presentan existe un aumento pero no tan considerable como en el caso anterior. La evaluación de los resultados, de estos experimentos, se realizó utilizando la técnica de *10-fold cross validation*. Además se realizó la prueba de significancia estadística con ambos métodos. La significancia estadística fue medida utilizando la prueba de *t-test* con un nivel de significancia de 0.05 y un intervalo de confidencia de 95 %. En este caso nuestro método mostró ser estadísticamente superior al método propuesto por [Bazell,2000].

6.5. Discusión

Para evaluar el método propuesto se creó la base de datos con placas digitalizadas de la colección del Instituto Nacional de Astrofísica Óptica y Electronica. Además, se realizaron diversas pruebas tanto para la clasificación plana como para la clasificación jerárquica. Por último, se llevaron a cabo las comparaciones entre los algoritmos de clasificación *Naive Bayes* y *Random Forest*. Los resultados obtenidos por el método propuesto se pueden analizar de la siguiente manera.

- **Separción de estrellas y galaxias**. En esta primera clasificación, el clasificador *Random Forest* fue el que presentó los mejores resultados y dado que solo se hizó una clasificación entre dos objetos no se vio necesaria la utilización de una clasificación jerárquica.

- **Clasificación plana de galaxias**. En esta clasificación se consideraron cuatro (E, Irr, S, S0) y nueve tipos (E, Irr, Sa, Sb, Sc, SBa, SBb, SBc, S0) de galaxias y el clasificador que obtuvo los mejores resultados en ambos casos fue *Random Forest*. Como se puede observar, los porcentajes de clasificación descienden al aumentar el número de clases dado existen clases que forman parte de otra como es el caso de las galaxias de tipo SBa, SBb y SBc ya que éstas pertenecen al tipo de espirales barradas, por lo que comparten varias de sus características y por lo tanto es dificil discernir entre ellas.

- **Clasificación jerárquica de galaxias**. Para esta clasificación se consideran nueve clases de acuerdo a la clasificación de Hubble (E, Irr, Sa, Sb, Sc, SBa, SBb, SBc, S0) y se utiliza el clasificador jerárquico multidimensional. La clasificación

jerárquica muestra una mejora significativa respecto a la clasificación plana, ya que presenta hasta un incremento del 14 % en la precisión de la clasificación. Además las pruebas nos muestran que la clasificación jerárquica se mejora si utilizamos tanto los ejemplos artificiales como la técnica de sobre-muestreo (*Ressampling*), pues cuando se utilizan por separado existe una mejora en los resultados.

A partir de los resultados obtenidos en estas pruebas, se demostró que la clasificación jerárquica de galaxias permite mejorar los niveles de clasificación comparado con un enfoque de clasificación plana, en donde no se puede explotar el uso de la relación que existe entre los diferentes tipos de galaxias. Además, las técnicas de *Resampling* y la creación de ejemplos artificiales mejoran la clasificación ya que ayudan a reducir el problema de desbalance de clases.

Capítulo 7

Conclusiones y Trabajo Futuro

7.1. Resumen

El Instituto Nacional de Astrofísica Óptica y Electrónica cuenta con una importante colección de placas astronómicas, ya que parte del estudio astronómico en nuestro país se llevó a cabo con las placas astronómicas tomadas en la Cámara Schmidt del INAOE. Estas placas comprenden el estudio del cielo por un periodo de aproximadamente 50 años, por lo que contienen gran cantidad de información histórica, por lo tanto el estudio de éstas es de gran importancia. Pero a pesar de la antigüedad de las placas, estas no han sido analizadas en su totalidad. De ahí el surgimiento del objetivo de este trabajo de tesis que consiste en diseñar e implementar un algoritmo capaz de segmentar y clasificar galaxias contenidas en placas astronómicas, de acuerdo a su morfología.

En esta tesis se propuso un método para la clasificación de galaxias en placas digitalizadas basado en un enfoque de clasificación jerárquica, el cual está dividido en siete etapas: En la primera y segunda se llevó a cabo el procesamiento de las imágenes (las placas digitalizadas), para ello se utilizaron métodos de segmentación y extracción de características. Se aplicó un método de umbralización para llevar a cabo la segmentación de cada una de las imágenes de la base datos, que se obtuvo a partir de imágenes de las placas digitalizadas, esto con el objetivo de identificar los objetos contenidos en la imagen. Las primeras características que fueron extraídas de cada uno de los objetos de las imágenes se obtuvieron a través de los momentos geométricos, los cuales permitieron

hacer una primera clasificación de los objetos contenidos en las placas astronómicas. En la tercera etapa se llevó a cabo la clasificación de estrellas y galaxias para así discernir entre estos objetos estelares. En la cuarta y quinta etapa se realizó la segmentación y extracción de características de las galaxias. En la segmentación se separó el núcleo de la galaxia del resto de cuerpo, esto se hizo utilizando los momentos geométricos y el radio de Petrosian. Esto con la finalidad de poder analizar tanto características del núcleo como de la parte restante de la galaxia de manera separada. Para la extracción de características se utilizó la lista de características del trabajo [Bazell et al., 2001]. En la sexta etapa se llevó a cabo la creación de ejemplos artificiales para resolver el problema de desbalance de clases. Finalmente, en la séptima etapa se realizó la clasificación de las galaxias utilizando un modelo jerárquico para realizar la clasificación de las galaxias. La jerarquía se basa en el diagrama de clasificación de Hubble, y consideramos un total de nueve tipos de galaxias, los cuales son: Sa, Sb, Sc, SBa, SBb, SBc, S0,I y E; éstos se ubican en las hojas de la jerarquía. La clasificación jerárquica se desarrolló bajo un enfoque de clasificación jerárquica multidimensional, utilizando la técnica de producto de probabilidades [Hernandez,2012].

Los experimentos se dividieron en dos partes principales, la primera corresponde a la clasificación plana y la segunda a la clasificación jerárquica. En la clasificación plana primero se consideraron cuatro tipos de galaxias (E, S, S0, Irr) y los clasificadores *Naive Bayes* y *Random Forest*. En esta primera prueba el clasificador que obtuvo mejores resultados fue *Random Forest* ya que obtuvo hasta un 64.17 % incluyendo ejemplos artificiales y *Resampling*. Después se consideraron nueve tipos de galaxias (E, Sa, Sb, Sc, SBa, SBb, SBc, S0, Irr) y se utilizaron los mismos clasificadores. Al igual que en el caso anterior, el clasificador que mostro mejores resultados fue *Random Forest* con un 39.44 %. Una vez realizada la clasificación plana se llevó a cabo la clasificación jerárquica utilizando un enfoque jerárquico multidimensional con la variante de producto de probabilidades y considerando nueve tipos de galaxias (E, Sa, Sb, Sc, SBa, SBb, SBc, S0, Irr). Se obtuvo una mejora considerable pues se paso de un 39.44 % a un 53.57 % con el clasificador *Random Forest*.

7.2. Conclusiones

Como primer paso para llevar a cabo la clasificación de las galaxias se realizó la separacion de estrellas y galaxias. Dado que los resultados obtenidos fueron buenos (más del 90 % de precisión) se consideran a los momentos geométricos como características que ayudan a discernir entre ambos objetos. Además el clasificador que mejores resultados presento durante los experimentos fue *Random Forest*.

Como resultado del trabajo realizado se concluye que la clasificación jerárquica muestra una mejora significativa respecto a a la clasificación plana, siendo la diferencia mayor mientras más clases se consideren. De acuerdo a los resultados obtenidos, la clasificación plana mostró una mejora al aplicar cualquiera de las dos técnicas que se utilizaron para el problema de desbalance de clases o bien al aplicar estas técnicas en conjunto. En el caso de la clasificación jerárquica sólo mostró esa mejoría al aplicar ambas técnicas ya que al crecer el conjunto de entrenamiento se puede establecer una mejor relación entre las clases. De las pruebas realizadas se demostró que el utilizar un enfoque de clasificación jerárquica es viable para utilizarse en sistemas de clasificación de galaxias.

7.3. Aportaciones

Las principales contribuciones de este trabajo son:

- Un método novedoso basado en un enfoque de clasificación jerárquica que comprende diversas etapas, para realizar la clasificación de galaxias en placas digitalizadas.
- Un método para atacar el problema de desbalance en la clasificación de galaxias utilizando ejemplos artificiales.
- Una nueva forma de llevar a cabo la separación del núcleo de la galaxia del resto del cuerpo, utilizando los momentos geométricos.

7.4. Trabajo Futuro

Como trabajo futuro se propone:

1. Incorporar un mayor número de placas a nuestra base de datos, para que ésta contenga un mayor número de ejemplos de galaxias y disminuir el problema de desbalance de clases.

2. Considerar el análisis y el uso de las isofotas (líneas de igual brillo que tiene un objeto, que perfilan los diferentes niveles de gris de la imagen) de las galaxias e incluirlas en la lista de características de las mismas.

3. Incluir aquellos tipos de galaxias que comprenden la transición entre un tipo y otro, ya que esto podría ayudar a tener más delimitados los tipos.

4. Combinar las características espectrales y morfológicas de las galaxias.

Bibliografía

[Abraham et al., 1994] Abraham R.G., Valdes F., Yee H. K. C.,ven den Bergh S.,The Morphologies of Distant Galaxies. I. An Automated Classification System, 1994, ApJ,432,75.

[Abraham et al.,2003] Abraham, R., van den Bergh S., and Nair, A New Approach to Galaxy Morphology. I. Analysis of the Sloan Digital Sky Survey Early Data Release P. 2003, ApJ, 588, 218

[Astikainen et al., 2008] Astikainen K, Holmand L, Pitkanen E, Szedmak S, Rousu J (2008) Towards structured output prediction of enzyme function. BMC Proc 2(Suppl 4)

[Bauer, 2008] T. Bauer,Astroinformatics - A Study About Constrains And Requirements For Next Generation Astronomical Image Processing. IADIS International Conference Informatics 2008.

[Bazell,2000] D. Bazell, Feature relevance in Morphological Galaxy Classification, Mon.Not.R. Astron. Soc. 316,519-528, 2000

[Bazell et al., 2001] D. Bazell and David W. Aha, Ensembles of Classifiers for Morphologycal Galaxy Classification. The Astrophysical Journal, 548:219-223, 2001 February.

[Bergh,1997] S. van den Bergh, The Evolution of Galaxies and Stellar Populations, B.M. Tinsley and R.B.Larson Eds., Yale Univ. Obs. 1977.

[Breiman, 2001] Breiman, L., Random forests. Machine Learning, pp 5-32,2001.

[Chong,2003] Chong Ho Yu, Resampling methods: concepts, applications, and justification. Practical Assessment, Research & Evaluation, 8(19),2003.

[Coenda, 2008] Valeria Coenda, Cúmulos de Galaxias: Propiedades de Galaxias y Subsistemas,Universidad Nacinal de Cordoba, 2008

[De la Calleja et al. 2004] Jorge de la Calleja and Olac Fuentes, Automated Classification of Galaxy Images. M.Gh. Negoita et al. (Eds): KES 2004, LNAI 3215, pp. 411-418, 2004, Springe-Verlag Berlin Heidelberg 2004.

[De la Calleja et al.,2010] Jorge de la Calleja, Gladis Huerta, Olac Fuentes, Antonio Benitez, The Imbalance Problem in Morphologica Galaxiy Classification, I. Bloch and R.M. Cesar, Jr.(eds); LNCS6419, pp. 533-540, 2010

[Diaz, 2005] Raquel Diaz Hernández, Análisis Espectrofotométrico de las Placas Astronómicas de la Cámara Schmidt de Tonantzintla, INAOE, 2005.

[Eskridge et al. 1999] Paul B. Eskridge and Jay A. Frogel,What is the true fraction of Barred Galaxies. Astrophysics and Space Science 269–270: 427–430, 1999.

[Freitas et al.,2007] Freitas AA, de Carvalho ACPLF, Research and trends in data mining technologies and applications, Idea Group, chap A: tutorial on hierarchical classification with applications in bioinformatics, pp 175–208, 2007.

[Hernandez,2012] Julio Noe Hernandez Torres, Clasificación Jerárquica Multidimensional, INAOE, 2012.

[Karttunen et al., 2007] H. Karttunen, P. Kröger, H. Oja, M. Poutanen and K. J. Donner (Eds.), Fundamental Astronomy. Springer, pp. 367-372, 2007

[Kohavi et al.,1995] Kohavi and Ron, A study of cross-validation and bootstrap for accuracy estimation and model selection, In IJCAI, pp 1137–1145, 1995.

[Koller et al.,1997] Koller D, Sahami M Hierarchically classifying documents using very few words. In: Proceedings of the 14th international conference on machine learning, pp 170–178, 1997.

[Lahav, 1996] Lahav O. Artificial neural networks as a tool for galaxy classification, in Data Analysis in Astronomy, Erice, Italy, 1996.

[Lotz et al., 2004] Jennifer M. Lotz, Joel Primack, Pero Madau, A New Nonparametric Approach to Galaxy Morphological Classification. The Astronomical Journal, 128:163-182, 2004.

[Martín, 2002] Marcos Martín, Descriptores de Imagen, UNAM, 2002.

[Michie et al., 1994] Michie D., Spiegelhalter D. J., Taylor C. C., Machine learning, neural and statistical classification, (edited collection). New York: Ellis Horwood, 1994.

[Mitchell, 1997] T. Mitchell. Machine Learning. McGraw-Hill, 1997.

[Morales et al.,2010] Eduardo F. Morales, Jesús A. González, El Problema de las Clases Desbalanceadas, INAOE, 2010

[Naim et al.,1995] Naim A., O. Lahav, R.J. Buta, H.G. Corwin, G. de Vaucouleurs, A Comparative Study of Morphologica Classification of APM galaxies 1995a, MNRAS, 274, 1107

[Petrosian,1982] Petrosian, A. R, On the connection between Seyfert galaxies and neighboring objects, Astrofizika, vol. 18, pp. 548-562, 1982

[Qiang Wu et al., 2002] Qiang Wu, Fatima A. Merchant, and Kenneth R. Castleman, Microscope image processing. Elsevier, 2008.

[Sanchez,2005] R. R. Sánchez, Selección de atributos mediante proyecciones, PhD thesis, Universidad de Sevilla, 2005.

[Shu,1982] Shu Frank H., The Physical Universe, An introduction to Astronomy. University Since Books, Sausalito, California, pp. 294, 1982.

[Simon et al. 1991] Simos J.,Bruce P., Resampling: A Tool for Everday Statistical Worf, Chance.4(1):22-31,1991.

[Skurichina et al., 2002] Skurichina M. and Duin R.P.W., Bagging, boosting and the random subspace method for linear classifiers. Pattern Analysis and Applications, 5(2):121–135, 2002.

[Stockman et al.,2001] G. Stockman and L. G. Shapiro. Computer Vision. Prentice-Hall, Upper Saddle River, NJ, USA, 2001.

[Tecpanecatl,2002] Tecpenecatl Mani Maria Saula Irma, Acervo de Placas Astronómicas del Instituto Nacional de Astrofísica Óptica y Electrónica, 2002.

[Ullman, 2001] Shimon Ullman, Erez Sali and Michel Vidal-Naquet, A Fragment-Based Approach to Object Representation and Classification, C. Arcelli et al.(Eds):IWVF4,LNCS2059,pp. 85-100,2001.

[Witten et al.,2005] Ian H. Witten, Eibe Frank, Data Mining, Practical Machine Learning Tools and Techniques, ELSEVIER, 2005

[Xiao et al.,2007] Xiao Z, Dellandréa E, Dou W, Chen L, Hierarchical Classification of Emotional Speech. Technical report RR-LIRIS-2007-006, LIRISUMR5205 CNRS/INSA de Lyon/Université Claude Bernard Lyon 1/Université Lumiére Lyon 2/Ecole Centrale de Lyon, 2007.

Apéndice A

Placas astronómicas

El proceso de etiquetado de las galaxias para la creación de la Base de Datos se llevo a cabo de manera manual y dicho proceso se describe en el diagrama de la figura A.1.

La Base de Datos se encuentra conformada por 24 imágenes de placas astronómicas que a su vez contienen 172 galaxias las cuales se encuentran distribuidas como se muestra en la tabla A.1. Además las galaxias de encuentran listadas en las tablas A.2 y A.3.

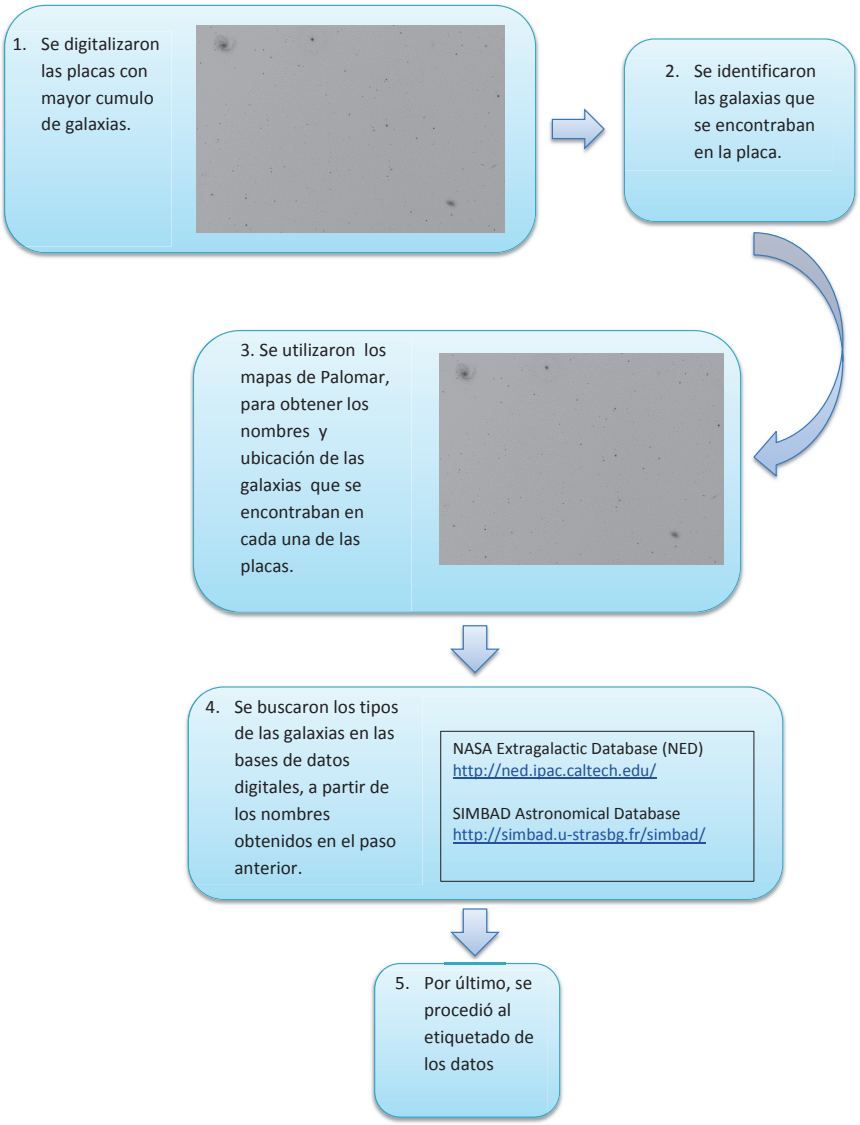

Figura A.1: Creación de la base de datos.

Nombre de placa	**Número de galaxias**
ST 242	14
ST 243	0
ST 248	2
ST 324	44
ST 386	8
AC 486	9
ST 882	11
ST 977	1
AC 3084	6
AC 3085	7
ST 4141	4
AC 7441	5
AC 8386	8
AC 8409	10
AC 8427	1
AC 8431	16
AC 8438	5
AC 8483	0
AC 8484	0
AC 8491	2
AC 8514	2
AC 8523	6
AC 8531	3
AC 8580	4

Tabla A.1: Relación del número de galaxias para cada una de las placas digitalizadas.

Nombre de placa	Nombre de galaxias	Nombre de placa	Nombre de galaxias
ST 242		ST 386	
	NGC5270, NGC5300		NGC5813, NGC5831
	NGC5335, NGC5338		NGC5838, NGC5846A
	NGC5348, NGC5356		NGC5850, NGC5854
	NGC5360, NGC5363		NGC5865, NGC5806
	NGC5364, UGC8534		
	UGC8686, UGC8774		
	UGC8818, NGC5252		
ST 243		AC 486	
			NGC5443, NGC5448
			NGC5448A, NGC5457
			NGC5473, NGC5475
			NGC5774, UGC08837
			NGC5422
ST 248		ST 882	
	NGC5661, NGC5457		NGC4136, NGC4150
			NGC4169, NGC4170
			NGC4175, NGC4185
			NGC4245, NGC4251
			NGC4274, NGC4134
ST 324		ST 977	
	IC3064, IC3392		NGC6221
	IC3476, NGC4189		
	NGC4192, NGC4212		
	NGC4216, NGC4222		
	NGC4237, NGC4267		
	NGC4294, NGC4298		
	NGC4299, NGC4302		
	NGC4305, NGC4306		
	NGC4313, NGC4330		
	NGC4351, NGC4371		
	NGC4374, NGC4387		
	NGC4402, NGC4406		
	NGC4407, NGC4419		
	NGC4425, NGC4429		
	NGC4435, NGC4438		
	NGC4452, NGC4458		
	NGC4459, NGC4461		
	NGC4473, NGC4477		
	NGC4486, NGC4491		
	NGC4497, NGC4501		
	NGC4503, UGC7345		
	UGC7365, 4474		
AC 3084		AC 3085	
	NGC3556, NEBU-NGC3587		NGC4288, NGC4346
	NGC3619, UGC6323		NGC4389, NGC4449
	UGC6444, UGC6458		NGC4460, UGC7690
			NGC4242

Tabla A.2: Relación de galaxias para cada una de las placas digitalizadas.

Nombre de placa	Nombre de galaxias	Nombre de placa	Nombre de galaxias
ST 4141		AC 7441	
	IC5041, NGC692 NGC699, IC5039		NGC3596, NGC3623 NGC3627, NGC3593 NGC3628
AC 8386		AC 8409	
	NGC5660, NGC5673 NGC5676, NGC5682 NGC5689, NGC5693 NGC5707, IC1029		NGC4274, NGC4278 NGC4283, NGC4310 NGC4314, NGC4393 NGC4414, NGC4448 NGC4559, NGC4251
AC 8427		AC 8431	
	AC8427		NGC4085, NGC4088 NGC4096, NGC4100 NGC4144, NGC4144 NGC4157, NGC4217 NGC4218, NGC4220 NGC4232, NGC4248 NGC4258, NGC4643 NGC4939, UGC7358 NGC3985
AC 3084		AC 3085	
	NGC3556, NGC3619 UGC6323, UGC6444 UGC6458, NEBU-NGC3587		NGC4288, NGC4346 NGC4389, NGC4449 NGC4460, UGC7690 NGC4242
ST 4141		AC 7441	
	IC5041, NGC692 NGC699, IC5039		NGC3596, NGC3623 NGC3627, NGC3593 NGC3628
AC 8386		AC 8409	
	NGC5660, NGC5673 NGC5676, NGC5682 NGC5689, NGC5693 NGC5707, IC1029		NGC4274, NGC4278 NGC4283, NGC4310 NGC4314, NGC4393 NGC4414, NGC4448 NGC4559, NGC4251
AC 8427		AC 8431	
	AC8427		NGC4085, NGC3985 NGC4088, NGC4096 NGC4100, NGC4144 NGC4144, NGC4157 NGC4217, NGC4218 NGC4220, NGC4232 NGC4248, NGC4258 NGC4643, NGC4939 UGC7358
AC 8438		AC 8491	
	NGC5879, NGC5907 NGC5908, UGC09797 NGC5867		NGC6181, NGC6168
AC 8514		AC 8523	
	NGC5859, NGC5857		NGC5982, NGC5985 NGC5987, NGC5989 NGC6015, NGC5981
AC 8531		AC 8580	
	NGC5740, UGC9499 NGC5690		NGC6340, NGC6434 UGC10713, IC1251

Tabla A.3: Relación de galaxias con cada una de las placas digitalizadas.

Apéndice B

Implementación del método

La implementación del método se llevo a cabo en el lenguanje C utilizando el software OpenCV (Open Source Computer Vision) el cual es una libreria de funciones para la visión de computadora en tiempo real. OpenCV es liberado bajo una licencia BSD, que es gratuito para uso académico y comercial. Cuenta con interfaces para los lenguajes de programació C, $C++$, Python y Java que se ejecutan tanto en Windows, Linux, Mac y Android. La libreria cuenta con mas de 2500 algoritmos optimizados. En nuestro caso se utilizó bajo es sistema operativo Linux. Las librerias de OpenCV se encuentran disponibles en: http://opencv.org/.

A continuación se describe la implementación de nuestro método a través del diagrama de bloques que se muestra en la figura B.1.

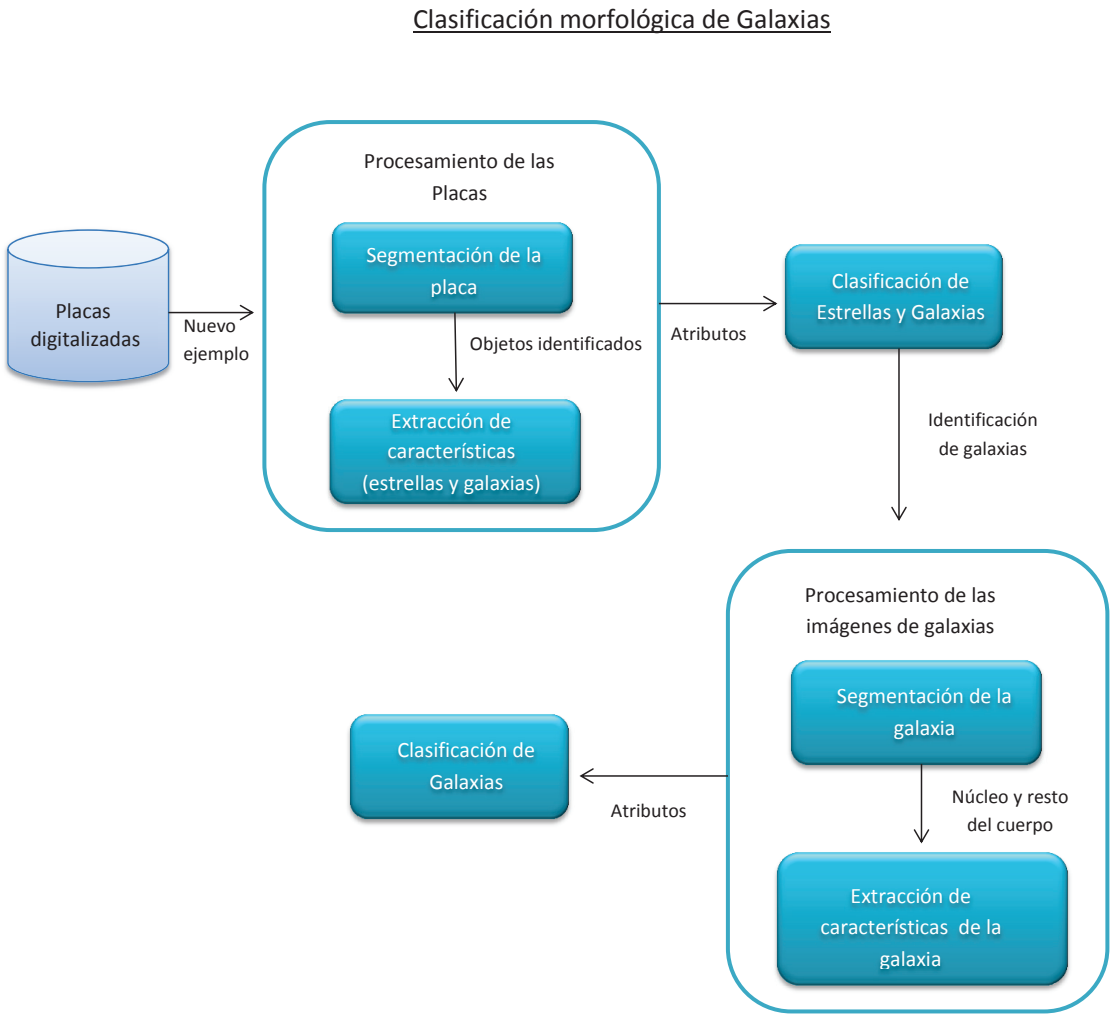

Figura B.1: Implementación del método.