Clasificación de imágenes de galaxias usando aprendizaje automático

Jorge de la Calleja, Antonio Benítez, Javier Caldera, Ma. Auxilio Medina
Ingeniería en Informática
Universidad Politécnica de Puebla
{jdelacalleja, abenitez, jcaldera, mmedina}@uppuebla.edu.mx

***Resumen*—** En este trabajo se presenta un método para clasificar imágenes de galaxias basándose en su morfología, aplicando métodos de aprendizaje automático. La morfología de las galaxias es generalmente un tema muy importante en el estudio del Universo, debido a que la clasificación de éstas se considera el primer paso que conduce a un mejor entendimiento tanto de su estructura como de su formación, y en consecuencia del Universo. El método propuesto está dividido en tres etapas: estandarización, compresión y clasificación de las imágenes, todo realizado en forma totalmente automatizada. Los resultados obtenidos, usando la técnica de *10-fold cross-validation*, muestran que el método permite clasificar correctamente a tres tipos principales de galaxias.

***Palabras Clave*— Aprendizaje automático, componentes principales, galaxias.**

2. Introducción

Desde la introducción del sistema de clasificación de *Hubble*, los astrónomos han buscado diferentes maneras para clasificar galaxias, esto porque es el primer paso hacia un mejor entendimiento del origen y los procesos de formación de las mismas, así como de la evolución del Universo. Esta tarea usualmente ha sido realizada de forma manual, observando de manera directa las placas fotográficas que han sido tomadas por telescopios de diversos programas astronómicos. Sin embargo, esta tarea no es tan sencilla debido a que se requiere de habilidad y experiencia, además de que consume mucho tiempo [6].

En años recientes, con numerosas exploraciones astronómicas realizadas en diferentes rangos de longitud de onda, la astronomía se ha convertido en una ciencia inmensamente rica en información. Por ejemplo, el *Sloan Digital Sky Survey* (http://www.sdss.org) producirá más de 50 millones de imágenes de galaxias, las cuales serán prácticamente imposibles de clasificar manualmente, y en consecuencia, se requerirán de métodos automáticos que ayuden con esta tarea. En la Figura 1 se muestra una imagen del espacio profundo, tomada por el telescopio espacial *Hubble* en 1996, aquí se pueden observar al menos 1,500 galaxias en varias etapas de su evolución.

Hay varias razones por las cuales la clasificación de galaxias es importante, así revisamos del porqué introducir métodos automáticos. La primera de ellas es porque se necesita producir grandes catálogos para obtener datos estadísticos y desarrollar programas observacionales. Otra razón es que los métodos automáticos son más imparciales que los humanos, esto es, no están sujetos a prejuicios tanto conscientes como inconscientes al momento de tomar la decisión de clasificar un objeto [1].

Figura 1. Imagen del espacio profundo. http://Hubblesite.org

Diversos estudios se han llevado a cabo en las últimas décadas para realizar clasificación automatizada de galaxias. Aquí sólo se presentan algunos de los más destacados. Por ejemplo, *Storrie-Lombardie et al*. [10], *Naim et al.* [6], *Lahav* [4], y Ball [1] usaron redes neuronales artificiales; *Owens* et al. [9] usaron árboles oblicuos de decisión; mientras que ensambles de clasificadores fueron empleados por *Bazell* y *Aha* [2]. *Odewahn et al.* [7] presentaron un enfoque basado en la transformada de *Fourier* para reconstruir imágenes de galaxias, mientras que *Godeyra* y *Lolling* [3] usaron dos tipos de clasificadores para realizar la clasificación.

En el enfoque presentado en este trabajo, se han aplicado tres métodos de aprendizaje automático: redes neuronales artificiales, regresión localmente ponderada y el clasificador de *Bayes*. Antes de realizar la clasificación de las imágenes de galaxias, éstas se han estandarizado (rotado, centrado y recortado) de forma totalmente automática. Además, se ha usado la técnica de análisis de

componentes principales (*Principal Component Analysis, PCA*) para reducir su dimensionalidad y para obtener atributos que las caractericen. Los resultados muestran que el método propuesto produce muy buenos resultados, particularmente para clasificar tres tipos de galaxias.

El resto del artículo está organizado de la siguiente manera: en la sección II se presentan conceptos relacionados con los diferentes temas que se abordan. En la sección III se presenta el método propuesto, describiendo las etapas que lo conforman. En la sección IV los resultados experimentales son mostrados y finalmente algunas conclusiones y trabajo futuro son presentados en la sección V.

3. Conceptos preliminares

II.1 Galaxias

Las galaxias son sistemas muy grandes formados de estrellas, polvo y nubes de gases, todos ellos unidos por la fuerza de gravedad [11]. Se puede decir que las galaxias son los bloques con los que se construye el Universo.

Antes de la mitad de la década de 1920, los astrónomos no estaban aún seguros si las galaxias eran sistemas separados de la nuestra, o si simplemente era otro tipo de nebulosa en nuestra galaxia. Sin embargo, en 1929, el astrónomo *Edwin Hubble* con su descubrimiento de que el Universo está en expansión, fue aceptado que las galaxias son como islas, tan grandes o más que la nuestra [1].

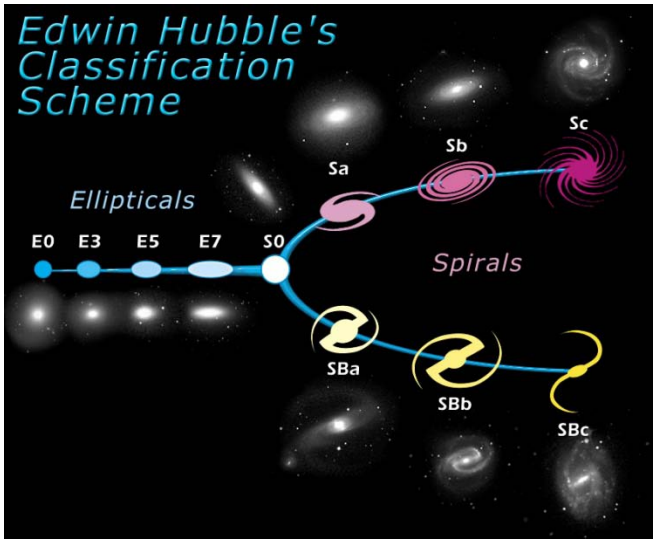

Figura 2. Esquema de clasificación de *Hubble*. http://www.sdss.org

Las galaxias tienen muchas características diferentes, sin embargo, la manera más fácil de clasificarlas es por su forma o morfología, y *Edwin Hubble* ideó un método básico para agruparlas (*Edwin Hubble's Classification Scheme*, Figura 2). En su esquema de clasificación, hay tres tipos principales de galaxias: Espirales, Elípticas e Irregulares. Las galaxias elípticas tienen la forma de una elipse, y normalmente contienen muy poca materia interestelar, que consiste de estrellas viejas [9].

Las galaxias espirales fueron las primeras en ser descubiertas, debido a que éstas son las más luminosas y cercanas a la nuestra. Las galaxias espirales se pueden dividir en ordinarias y con barra. Las primeras tienen un núcleo aproximadamente esférico, mientras que las segundas tienen un núcleo que asemeja a una barra. Así también, las galaxias espirales pueden ser sub-clasificadas en Sa, Sb, Sc, SBa, SBb y SBc, dependiendo de su núcleo y proximidad de sus brazos hacia el núcleo. Una galaxia Sa tiene un núcleo más grande que una Sc, pero los brazos de una Sc están más separados que los de una Sa [13]. Nosotros vivimos en una galaxia de tipo espiral, llamada la *vía láctea*.

Por último, las galaxias irregulares no tienen una forma obvia elíptica o espiral; cuya forma la obtienen por la distorsión causada por la fuerza de gravedad de sus vecinos intergalácticos [9].

II.2 Aprendizaje automático

Aprender, es probablemente el rasgo más distintivo que tenemos los seres humanos. Esto incluye, la adquisición de información, transformarla en conocimiento, desarrollo de habilidades, organización de ésta, descubrimiento de nuevos hechos, entre otros muchos procesos.

El aprendizaje automático (*Machine learning*) intenta imitar estos procesos al estudiarlos y modelarlos de manera computacional.

Tom Mitchell, investigador de la Universidad de *Carnegie Mellon* de los Estados Unidos, establece que "… un programa de computadora se dice que *aprende* de la experiencia tomada al realizar una clase de tarea midiendo su desempeño, siempre y cuando este desempeño sobre la tarea efectuada pueda mejorarse por medio de la experiencia que va obteniendo…". Por ejemplo, un programa de computadora que aprende a jugar rompecabezas podría mejorar su desempeño a través de la experiencia que obtenga al jugar varios juegos contra expertos humanos o incluso contra él mismo.

El aprendizaje automático forma parte de un área muy amplia llamada inteligencia artificial (*Artificial intelligence*), y además se relaciona estrechamente con disciplinas como la visión por computadora, la robótica,

la probabilidad, la estadística, la biología, la teoría de información, entre muchas otras.

II.3 Algoritmos de aprendizaje automático

En esta sección se da una muy breve descripción de los métodos usados, por lo que se recomienda al lector revise las referencias citadas, y en particular [5].

Redes neuronales artificiales

Este método de aprendizaje automático es y ha sido uno de los más utilizados en diversas aplicaciones. Las redes neuronales artificiales (*Artificial neural networks*) han sido inspiradas por la observación de los sistemas biológicos neuronales que están formadas por conjuntos de unidades llamadas neuronas que se encuentran densamente interconectadas.

Computacionalmente hablando, existen varias topologías para crear redes neuronales, sin embargo, las más comunes son las redes *feed forward* y las *recurrent networks*. El primer tipo tiene los nodos (neuronas) organizados en una serie de capas llamadas de entrada, ocultas y de salida. En el segundo tipo de red se pueden formar topologías arbitrarias.

Una red neuronal funciona de la siguiente manera: cada nodo produce un valor de salida, que resulta de combinar los valores de entrada de los nodos que le anteceden, de esta forma se alimenta a los siguientes nodos de las capas siguientes hasta producir un valor final en la capa de salida.

Independientemente de la topología de una red neuronal, el primer paso es su entrenamiento, donde los pesos (valores) de los nodos son determinados. Uno de los algoritmos más utilizados para realizar esta etapa de entrenamiento es el llamado algoritmo de retro-propagación (*Backpropagation*) [5].

Regresión localmente ponderada

El algoritmo de regresión localmente ponderada (*Locally weighted regression, LWR*) pertenece a la familia de métodos basados en instancias. La idea básica de estos métodos consiste en almacenar todos los datos de entrenamiento disponibles, y cuando un dato nuevo se requiere clasificar, se buscan los datos más parecidos a éste para poder asignarle el valor correspondiente [5]. Para este trabajo se usó una función lineal, por lo que se dice que la regresión es lineal localmente ponderada (*Locally weigthed lineal regression*).

Clasificador de Bayes

El clasificador de *Bayes* es un algoritmo probabilístico que se basa en la suposición de que todos los valores de los atributos del conjunto de datos son condicionalmente independientes para los valores objetivo. El clasificador de *Bayes* se utiliza en tareas de aprendizaje en donde cada dato puede ser descrito como una *tupla* de valores-atributo y la función objetivo puede tomar cualquier valor de un conjunto finito de valores. Para clasificar un dato nuevo, este algoritmo asigna la probabilidad más alta de acuerdo con la función objetivo [5].

II.4 Análisis de componentes principales

El análisis de componentes principales (*Principal component analysis, PCA*) es un método estadístico que puede ser usado tanto para comprimir datos como para encontrar información relevante en los mismos. En algunas áreas, *PCA* es también llamado la transformación de *Karhunen-Loève* o de *Hotelling*.

Este método busca un conjunto de vectores perpendiculares y sus valores asociados que mejor describen a la distribución de un conjunto de datos [12]. Los vectores y valores calculados son los *eigenvectores* y *eigenvalores*, respectivamente, de la matriz de covarianza. Los *eigenvalores* asociados permiten ordenar a los *eigenvectores* de acuerdo a su utilidad para caracterizar la variación entre los datos. El primer *eigenvector* (componente principal) es el que proporciona la máxima varianza en el conjunto de datos.

4. El método propuesto

El método que se propone en este trabajo para clasificar imágenes de galaxias de forma totalmente automatizada, está dividido en tres partes: Estandarización, Compresión y Clasificación de las imágenes.

El método funciona de la siguiente manera: Primero, las imágenes son rotadas, centradas, recortadas y estandarizadas a un tamaño determinado, de forma totalmente automática. Enseguida, estas imágenes estandarizadas son reducidas en su dimensionalidad y se busca un conjunto de características (componentes principales) que permitan diferenciarlas. La proyección de las imágenes sobre los componentes principales serán los parámetros (atributos) para realizar la última parte, la de clasificación. Las siguientes tres sub-secciones describen con un poco más de detalle cada etapa del método.

III.1 Estandarización de imágenes

El objetivo de este primer paso es crear imágenes invariantes al color, posición, orientación y tamaño; esto porque las imágenes de galaxias generalmente están en diferentes tamaños y formatos de color, así también porque es muy común que la galaxia contenida en la imagen no se encuentra en el centro de la misma.

Por lo tanto, primero se localiza la galaxia contenida en la imagen aplicando un *threshold* (umbralización).

Después se calcula el centro de la imagen dado por la columna y fila correspondiente. Enseguida, se calcula la matriz de covarianza de los puntos en la imagen. El eje principal de la galaxia esta dado por el primer *eigenvector* (el que tiene el *eigenvalor* más grande) de la matriz de covarianza. Así, se rota la imagen de tal manera que el eje principal de la galaxia quede horizontal. Después, se centra la galaxia y se recorta la imagen, eliminando las columnas y filas que sólo contienen pixeles del color del fondo (generalmente de color negro). Por último, las imágenes son ajustadas a un tamaño de 128x128 pixeles. La Figura 3 muestra algunos ejemplos del procesamiento para diferentes tipos de galaxias.

III.2 Compresión de imágenes

La idea general para realizar clasificación de objetos, y particularmente de imágenes de galaxias, es primero extraer información relevante de los objetos, codificarla lo mejor posible y comparar algún objeto nuevo con esta información ya codificada para asignarle su clasificación. Por lo tanto, para realizar esta tarea de búsqueda y extracción de información relevante, se ha usado el análisis de componentes principales. Y como se explicó anteriormente, la idea básica de *PCA* es buscar los *eigenvectores* de la matriz de covarianza del conjunto de imágenes de galaxias, de tal manera que proporcionen la máxima cantidad de varianza entre ellas. Así, estos *eigenvectores* pueden ser vistos como el conjunto de características que permitirán diferenciar cada tipo de galaxia en el conjunto.

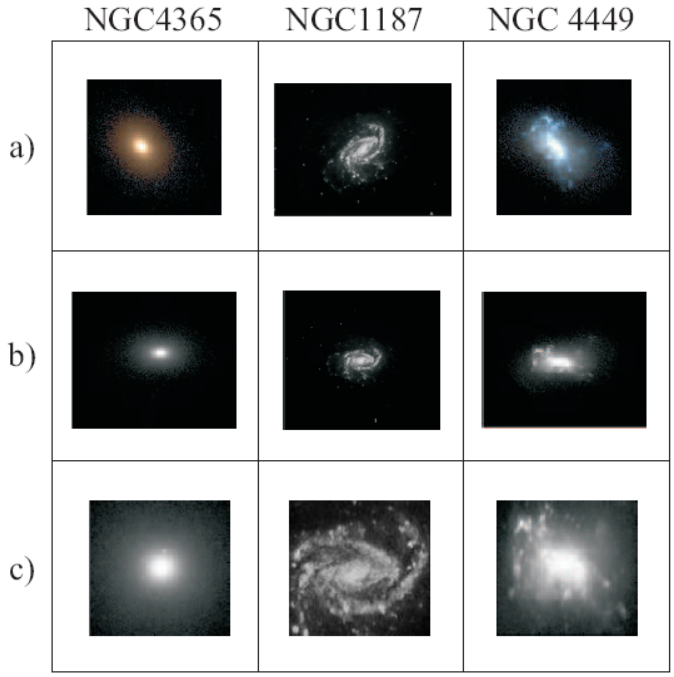

Figura 3. Ejemplo del procesamiento de imágenes de galaxias: a) Imágenes originales, b) Imágenes rotadas, y c) Imágenes recortadas y centradas.

Las imágenes estandarizadas obtenidas en la etapa previa, son de tamaño 128x128 pixeles, y describen un vector cuya dimensión es de 16,384 puntos. Por lo que se tendrá una matriz muy grande de tamaño de 16,384 (columnas) por el número de imágenes de galaxias (renglones). Sin embargo, al reducir este conjunto de datos usando *PCA*, se puede obtener una matriz mucho más pequeña, de tamaño: número de componentes principales (columnas) por número de imágenes (renglones).

III.3 Clasificación de imágenes

Una vez que se ha reducido la dimensionalidad del conjunto de datos y se han encontrado parámetros de entrada para los algoritmos de clasificación, se puede realizar la tarea de clasificación. Para ello, se han aplicado los algoritmos de redes neuronales artificiales, regresión lineal localmente ponderada y el clasificador de *Bayes*.

La etapa de clasificación comprende dos pasos. En el primero, los algoritmos de aprendizaje son entrenados, de acuerdo a lo presentado en secciones anteriores, usando un conjunto de imágenes llamado de entrenamiento. Enseguida, otro conjunto más pequeño de imágenes es proporcionado a los algoritmos como conjunto de prueba, de tal manera que sobre este último se pueda verificar la correcta clasificación de las imágenes.

5. Resultados experimentales

IV.1 Conjunto de datos

El conjunto original de datos consistió de 292 imágenes de galaxias, la mayoría tomadas del catálogo de internet de la *Astronomical Society of the Pacific* (http://www.astrosociety.org), y su clasificación del catálogo en línea de *Students for the Exploration and Development of Space* (*SEDS*) [10]. Obviamente el conjunto de datos en Internet es mucho más grande que 292 imágenes; sin embargo, sólo se tomó un pequeño conjunto de estas para probar la efectividad del método propuesto. Así también, para nuestro fin, sólo se consideraron tres, cinco y siete tipos de galaxias de acuerdo con el esquema de clasificación de *Hubble*.

IV.2 Parámetros de entrada

Como parámetros de entrada para los algoritmos de aprendizaje automático se probaron diferentes números de componentes principales, dependiendo de la información que representaban. Para este trabajo sólo se consideraron trece y veinticinco componentes principales, que representan el 80% y 90%, respectivamente, de la información contenida en el conjunto original de imágenes de galaxias.

Así, de la matriz original de 16,384 columnas por 292 renglones, se tienen matrices de 13x292 y de 25x292.

IV.3 Resultados

Para realizar los experimentos se ha usado una técnica llamada validación cruzada (*cross-validation*), en particular se usó *10-fold cross-validation*. Esta técnica consiste en dividir el conjunto de datos original en diez partes iguales, de las cuales nueve partes son tomadas como conjunto de entrenamiento y una parte como conjunto de prueba. Así, cada experimento se repite diez veces de tal manera que todo el conjunto de datos sea utilizado como entrenamiento y prueba. Los resultados que a continuación se presentan corresponden al promedio de realizar cinco ejecuciones de *10-fold cross-validation* para cada algoritmo.

Para realizar los experimentos con redes neuronales artificiales se usó el *feedforward neural network* que está implementado en *Matlab Neural Network Toolbox™*. La red usada tuvo diferentes configuraciones dependiendo del número de parámetros de entrada, del número de nodos para la capa oculta y para la capa de salida. Todas las redes fueron entrenadas usando el algoritmo de *backpropagation* durante 100 épocas y con un coeficiente de aprendizaje de 0.0015.

El algoritmo de regresión localmente ponderada fue implementado en *Matlab™*, usando un modelo lineal y diferentes números de puntos para realizar la aproximación. Así también se ponderó la contribución de cada dato usando la distancia *Euclidiana*.

Para el caso del clasificador de *Bayes*, se utilizó la implementación de *Weka* (*Waikato Environment for Knowledge Analysis, http://www.cs.waikato.ac.nz/ml/weka/*), que es un ambiente de libre distribución que contiene varios algoritmos de aprendizaje automático.

La Tabla 1 muestra los resultados obtenidos para cada algoritmo, considerando 3, 5 y 7 tipos (clases) de galaxias. Como se puede observar, el algoritmo de regresión lineal localmente ponderada fue el que obtuvo los mejores resultados, con 91.50% de exactitud para tres clases, 50.35% de exactitud para cinco clases y 43.63% de exactitud para siete clases. Así también, se puede destacar que usando 13 atributos (componentes principales) son suficientes como parámetros de entrada para los clasificadores, obteniendo los mejores resultados.

Redes neuronales artificiales			
#Atributos	**3 clases**	**5 clases**	**7 clases**
13	88.50	49.00	43.50
25	85.95	43.07	37.79
Regresión lineal localmente ponderada			
13	**91.50**	**50.35**	43.59
25	88.92	49.92	**43.63**
Bayes			
13	80.68	44.53	37.46
25	75.88	40.33	34.58

Tabla 1. Exactitud obtenida por los diferentes algoritmos. En negrita se muestran los mejores resultados para cada clase.

6. Conclusiones

Se ha presentado un método que permite clasificar imágenes de galaxias de forma totalmente automática. El método utiliza los algoritmos de redes neuronales, regresión lineal localmente pondera y el clasificador de *Bayes*. Así también, se utilizó la técnica de análisis de componentes principales para reducir el conjunto de datos y obtener atributos para diferenciar a las galaxias. Los resultados obtenidos son muy buenos, y en algunos casos, son mejores que los presentados en la literatura. El trabajo futuro incluye probar otros algoritmos de aprendizaje automático, así como experimentar con otros métodos para reducir y obtener atributos de las imágenes.

REFERENCIAS

[1] Ball, N. Morphological classification of galaxies using artificial neural networks. Master's thesis, University of Sussex, 2002.

[2] Bazell, D. and Aha, D. Ensembles of classifiers for morphological galaxy classification, The Astrophysical Journal, 548:219-223, 2001.

[3] Goderya, S. and Lolling, S. Morphological classification of galaxies using computer vision and anns. Astrophysics and Space Science, 279(377), 2002.

[4] Lahav, O. Galaxy classification by human eyes and by artificial neural networks. The Astrophysical Journal, 1996.

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

[6] Naim, A., O, O. L., Sodré, J., and Storrie-Lombardi, M. Automated morphological classification of apm galaxies by supervised artificial neural networks. Monthly Notices of the Royal Astronomical Society, 275(567), 1995.

[7] Odewahn, S., Cohen, S., Windhorst, R., and Philip, N. Automated galaxy morphology: a fourier approach. The Astrophysical Journal, 568:539-557, 2002.

[8] Owens, E., Griffiths, R., and K.U., K. R. Using oblique decision trees for the morphological classification of galaxies. Monthly Notices of the Royal Astronomical Society, 281(153), 1996.

[9] Students for the Exploration and Development of Space (SEDS). http://seds.org. 2011

[10] Storrie-Lombardi, M., O., O. L., Sodré, L., and Storrie-Lombardi, L. Morphological classification of galaxies by artificial neural networks. Monthly Notices of the Royal Astronomical Society, 259(8), 1992.

[11] The Anglo Australian Observatory. Galaxies near and far. http://www.ast.cam.ac.uk. (2011)

[12] Turk, M. and Pentland, A. Face recognition using eigenfaces. In Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, 1991.

[13] University of Washington, Department of Astronomy. Hubble galaxy classification. http://www.astro.washington.edu/labs/clearinghouse/labs/Hubclass/hubbleclass.html. 2011.