ISSN 1870-4069

Clasificación de galaxias utilizando procesamiento digital de imágenes y redes neuronales artificiales

Ricardo Cordero-Chan[1], Mauricio Gabriel Orozco-del-Castillo[1], Mario Renan Moreno-Sabido[1], Jorge Javier Hernández-Gómez[2], Gerardo Cetzal-Balam[1], Carlos Couder-Castañeda[2]

[1] Instituto Tecnológico de Mérida, Departamento de Sistemas y Computación, Yucatán, México

[2] Instituto Politécnico Nacional, Centro de Desarrollo Aeroespacial, Ciudad de México, México

xxricardo992xx@hotmail.com, mauricio.orozco@itmerida.edu.mx, {xacdc12,gerardoce23}@gmail.com, jjhernandezgo,ccouder}@ipn.mx

Resumen. El estudio de la formación y la evolución de las galaxias requiere de la medición de sus parámetros morfológicos. Tradicionalmente, el análisis morfológico se ha llevado a cabo principalmente a través de la extracción y selección de características, o mediante la inspección visual de expertos en un proceso que consume muchos recursos y que resulta prácticamente imposible de realizar en colecciones masivas de imágenes. A pesar de que se han realizado intentos para construir sistemas de clasificación automatizados, estos aún no tienen el nivel deseado de precisión que se requiere. En este trabajo se desarrolla una red neuronal convolucional, entrenada con la base de datos masiva del proyecto Galaxy Zoo, capaz de ser aplicada para la clasificación automática de la morfología de galaxias en imágenes. Para aumentar la precisión en la clasificación de la red neuronal, se preprocesan las imágenes de entrenamiento mediante la técnica de análisis de componentes principales. Este enfoque puede ser fundamental para el análisis y clasificación de imágenes de bases de datos mayores provenientes de proyectos aún en desarrollo, pues reduce la carga de trabajo de los científicos expertos y no depende de la interpretación manual inexperta de las imágenes.

Palabras clave: análisis de datos, predicción y clasificación, aprendizaje de máquinas, red neuronal artificial convolucional, analisis de componentes principales, ACP, RNA, inteligencia artificial, reconocimiento de patrones, clasificación de galaxias, morfología de galaxias.

Digital Imaging Processing and Artificial Neural Networks for Galaxy Classification

Abstract. The study of the formation and evolution of galaxies requires the measurement of their morphological parameters. Traditionally,

morphological analyses have been performed through the extraction and selection of features, or through visual inspection by experts in a high-burden process which is almost impossible to perform in massive image collections. Although there have been several attempts to build automated classification systems, these do not possess the required precision level. In this work we developed a convolutional artificial neural network and trained it with the massive database of the Galaxy Zoo project. This neural network can be applied to the automatic classification of images of galaxies according to their morphology. To increase the precision in the classification of the neural network, the training images are preprocessed using a principal component analysis approach. By reducing the work burden of experts and by not depending on the inexpert manual interpretation of images, this scope could be fundamental for the analysis and classification of images coming from even wider surveys which are currently under development.

Keywords: data analysis, prediction and classification, convolutional artificial neural network, machine clearning, principal component analysis, PCA, ANN, artificial intelligence, pattern recognition, galaxy classification, galaxy morphology, AI.

1. Introducción

Las galaxias muestran una gran variedad de aspectos morfológicos como formas, tamaños, colores, etcétera. Estas propiedades son importantes indicadores de su edad, su proceso de formación, así como de potenciales interacciones históricas con otros cuerpos celestes. Los estudios de formación y evolución de galaxias utilizan su morfología para evaluar los procesos físicos que les dan origen, sin embargo, requieren de la observación de un gran número de galaxias y la clasificación precisa de sus morfologías. La morfología de una galaxia puede ser derivada tanto por parámetros morfológicos, tales como concentración, asimetría, el coeficiente de Gini, etc. [3], así como por la inspección visual de imágenes de galaxias [15]. El enfoque visual es generalmente más resistente a los cambios en la resolución de señal-ruido en imágenes [14], lo que lo hace un método ideal para determinar la morfología de una galaxia. La clasificación de galaxias en categorías basadas en su morfología ha sido una práctica estándar desde que fue sistemáticamente aplicada por Hubble [9], y ha mostrado ser un aspecto muy relevante para su clasificación y posterior estudio. Por un lado, es un rastreador de la dinámica orbital de las estrellas en ella, pero también implica una huella de los procesos que impulsan la formación de estrellas y la actividad nuclear en las galaxias. La morfología visual de las galaxias también produce clasificaciones que están fuertemente correlacionadas con otros parámetros físicos. Por ejemplo, la presencia de múltiples núcleos y las características de las mareas extendidas parecen indicar que el mecanismo dominante que impulsa la formación de estrellas es una fusión en curso. De la misma forma, la ausencia de tales características podría implicar que la evolución de la galaxia puede estar siendo impulsada por procesos más lentos [1]. Inclusive, se ha sugerido

que la morfología de las galaxias puede proveer señales de su contenido de materia obscura, lo que constituye una prueba fehaciente del modelo cosmológico ΛCDM, permitiendo calibrar el contenido de materia ordinaria, materia y energía obscuras del universo [2,7,12,13].

Existen estudios a gran escala del espacio, tales como el Sloan Digital Sky Survey (SDSS, o Estudio Celeste Digital de Sloan). El SDSS es un estudio de una gran parte del cielo del norte que provee fotometría en cinco filtros: u, g, r, i y z [3]. El estudio cubre aproximadamente el 26 % del cielo completo. Estudios como el SDSS han resultado en la disponibilidad de datos de millones de objetos celestes en forma de imágenes, pero su análisis resulta prohibitivo para investigadores individuales o incluso para equipos de trabajo.

Se han realizado intentos para desarrollar sistemas de clasificación automática de la morfología de las galaxias, pero ha sido muy difícil alcanzar los niveles de confiabilidad requeridos para el análisis científico [2]. Recientemente fue diseñado y lanzado públicamente el proyecto Galaxy Zoo [19, 20], un proyecto concebido para acelerar esta tarea de clasificación que consiste en un método novedoso para desarrollar clasificaciones visuales a gran escala de conjuntos de datos. Utilizando más de medio millón de voluntarios, el proyecto ha clasificado, mediante la inspección visual directa, la muestra espectroscópica del SDSS. Con más de 40 clasificaciones por objeto, el proyecto Galaxy Zoo provee tanto una clasificación visual y una incertidumbre asociada (misma que sería muy complicada de estimar con un pequeño grupo de clasificadores humanos). Se ha comprobado que las clasificaciones del proyecto tienen una precisión comparable con aquellas derivadas por astrónomos expertos [13].

Actualmente existen estudios fotométricos de gran escala con el objetivo de recolectar datos para cientos de millones e incluso billones de estrellas y galaxias. Debido al gran volumen de datos, no es posible para los expertos humanos clasificarlos manualmente, y la separación de catálogos fotométricos en estrellas y galaxias tiene que ser automatizada. Casi todos los clasificadores de galaxias publicados en la literatura utilizan la limitada información disponible de catálogos astronómicos. Construir catálogos requiere de experiencia considerable en el campo para transformar los valores de los pixeles que representan a una imagen en características adecuadas, tales como magnitudes o información de la forma de un objeto.

Utilizando inteligencia artificial (IA), es posible utilizar algoritmos para crear automáticamente clasificación de distintas estructuras. En la rama del aprendizaje de máquinas (*machine learning*) llamada aprendizaje profundo (*deep learning*) [12], las características no son diseñadas por expertos humanos, sino que son aprendidas directamente de la información por Redes Neuronales Artificiales (RNAs). Los métodos de aprendizaje profundo aprenden múltiples niveles de características al transformar la característica en un nivel en una característica más abstracta en un nivel más alto. Estas múltiples capas de abstracción amplifican progresivamente los aspectos de las entradas de la red que son importantes para tareas de clasificación. En los últimos años, las técnicas de IA han adquirido mucha popularidad en distintas áreas de la astronomía [3, 4,

8]. Las RNAs fueron utilizadas por primera ocasión al problema de clasificación de galaxias en 1992 [16], y se han convertido en una parte fundamental de la astronomía [10]. Otros ejemplos exitosos al aplicar técnicas de IA al problema de la clasificación de galaxias incluyen los árboles de decisión [19], máquinas de vectores de soporte [5] y estrategias de combinación de clasificadores [11].

Por otro lado, el Análisis de Componentes Principales (ACP) tiene sus antecedentes en psicología, a través de las técnicas de regresión lineal iniciadas por Galton [7]. El nombre de "componentes principales" y su primer desarrollo teórico se deben a Hotteling [8], quién desarrolló un método de extracción de factores.

El ACP es una técnica de análisis estadístico multivariable que se clasifica entre los métodos de simplificación o reducción de la dimensionalidad de variables, y que se aplica cuando se dispone de un conjunto elevado de variables con datos cuantitativos y con el fin de obtener un conjunto menor de ellas, las componentes principales, que son una combinación lineal de las variables originales. Cuando las variables originales están muy correlacionadas entre sí, la mayor parte de su variabilidad se puede explicar con muy pocos componentes. Si las variables originales estuvieran completamente no correlacionadas entre sí, entonces el ACP carecería de aplicación.

Desde cierto punto de vista, el ACP permite identificar patrones en un conjunto de datos y expresar a estos de manera que sus similitudes y diferencias puedan ser recopiladas. De las primeras aplicaciones científicas del ACP, Turk y Pentland [18] lo aplicaron como una técnica de reconocimiento de patrones en el reconocimiento de rostros. Ellos enfocaron su investigación hacia desarrollar un modelo de reconocimiento de patrones que no dependiera de disponer de información tridimensional o geometría detallada. También ha sido exitosamente utilizado en múltiples campos como el reconocimiento de geocuerpos en datos sísmicos [17].

En este trabajo se propone un sistema basado en una RNA convolucional para la clasificación morfológica de galaxias. Utilizando imágenes de galaxias obtenidas directamente de la base de datos del SDSS, aplicamos técnicas de procesamiento digital de imágenes (PDI) para enfatizar características de interés en ellas, para después relacionarlas con las clasificaciones realizadas en las bases de datos publicadas por el proyecto Galaxy Zoo. En este documento se presentan los avances realizados hasta el momento en este trabajo, que actualmente se encuentra aún en proceso.

A la fecha se cuenta con un conjunto pequeño de imágenes de entrenamiento y se ha limitado el problema a la clasificación de galaxias en tres tipos, 1) elípticas, 2) espirales e 3) inciertas, sin embargo, debido a la comparable precisión con lo reportado por el proyecto Galaxy Zoo, los resultados preliminares son muy alentadores. Este artículo se organiza como sigue: en la Sección 2 se presenta tanto la metodología seguida como los resultados, presentando los datos utilizados (Sección 2.1) así como los métodos a utilizar (Secciones 2.2 y 2.3), mientras que en la Sección 3 se presentan algunas conclusiones sobre este estudio. Finalmente, en la Sección 4 se presenta el trabajo a seguir para concluir esta investigación.

2. Metodología y resultados

2.1. Los datos de Galaxy Zoo

Galaxy Zoo es un proyecto donde a los usuarios se les pide describir la morfología de galaxias basándose en imágenes a color [19, 20]. A los participantes se les hacen varias preguntas tales como "¿qué tan redonda es la galaxia?" y "¿tiene una acumulación central?", donde las respuestas de los usuarios determinan qué pregunta se realizará a continuación. Cuando muchos participantes han clasificado la misma imagen, sus respuestas se agregan en un conjunto de fracciones de votos ponderados. Estas fracciones de votos se utilizan para estimar niveles de confianza para cada respuesta, y son indicativas de la dificultad que los usuarios experimentaron al clasificar la imagen.

Más de medio millón de personas han contribuido clasificaciones a Galaxy Zoo, con cada imagen siendo clasificada por 40 a 50 personas [4]. Los datos del proyecto Galaxy Zoo han sido utilizados en una gran variedad de estudios de estructura, evolución y formación de galaxias [19, 22, 28]. Las comparaciones de las morfologías reportadas por este proyecto con muestras más pequeñas, tanto de expertos como de clasificaciones automáticas, muestran altos niveles de concordancia, testificando la precisión de las anotaciones masivas de los voluntarios de Galaxy Zoo.

Las imágenes de galaxias y los datos morfológicos fueron obtenidos de la SDSS utilizando los datos publicados por Galaxy Zoo. Un extracto de los datos publicados se muestra en la Tabla 1. Esta tabla contiene los datos de todas las galaxias de la Muestra de las Principales Galaxias (MGS, por sus siglas en inglés, Main Galaxy Sample), es decir, 667,945 galaxias. La tabla original incluye un identificador de cada objeto (ObjID), las coordenadas en formato Longitud o Ascención Recta del nodo (RA, por sus siglas en inglés, *Right Ascension*) y Dec (*Declination*), los votos crudos (N), los votos ponderados en categorías elípticas (E), galaxias espirales en el sentido de las manecillas del reloj (CW), en el sentido contrario (ACW), espirales no incluidas en las categorías anteriores (B), no sabe (DK), fusión de galaxias (MG), espirales combinadas (CS), y sin sesgo en dos categorías, elípticas (E) y espirales combinadas (CS), y banderas indicando la inclusión de la galaxia en un catálogo sin sesgo, espirales (S), elípticas (E) e inciertas (I). En este trabajo hacemos referencia únicamente a las columnas ObjID, RA, Dec, N, S, E e I.

De los datos publicados por el proyecto Galaxy Zoo [6], utilizamos aquellos correspondientes a las coordenadas de la galaxia (RA y Dec), y las correspondientes banderas que clasifican a la galaxia en una de tres categorías: espirales (S), elípticas (E) e inciertas (I). Con esta información, se construyó una base de datos con la cual fue posible descargar las imágenes correspondientes a las coordenadas reportadas por Galaxy Zoo del sitio web de SDSS. Hasta el momento se han recolectado imágenes correspondientes a 504 galaxias, tanto espirales como elípticas e inciertas. Un subconjunto de estas imágenes se muestra en la Fig. 1.

Tabla 1. Extracto de la tabla que muestra la clasificación de galaxias [6]. La tabla incluye un identificador de cada objeto (ObjID), las coordenadas en formato RA (*Right Ascension*) y Dec (*Declination*), los votos crudos (N) y banderas indicando la inclusión de la galaxia en un catálogo sin sesgo, espirales (S), Elípticas (E) e inciertas (I).

	Coordenadas			Banderas		
ObjID	RA	Dec	N	S	E	I
587727178986356000	00:00.4	-10:22:25.7	59	0	0	1
587727227300741000	00:00.7	-09:13:20.2	18	1	0	0
587727225153257000	00:01.0	-10:56:48.0	68	0	0	1
587730774962536000	00:01.4	+15:30:35.3	52	0	1	0
587731186203885000	00:01.6	-00:05:33.3	59	0	0	1
587727180060098000	00:01.6	-09:29:40.3	28	0	0	1
587731187277627000	00:01.9	+00:43:09.3	38	0	0	1
587727223024189000	00:02.0	+15:41:49.8	26	1	0	0
587730775499407000	00:02.1	+15:52:54.2	62	0	0	1
587727221950382000	00:02.4	+14:49:19.0	31	1	0	0
587730774425665000	00:02.6	+15:02:28.3	24	0	0	1

2.2. Análisis de componentes principales

Antes de entrenar a la RNA, estas imágenes de entrenamiento son sujetas a un proceso basado en ACP. El sistema aquí desarrollado tiene como base el trabajo de Turk y Pentland [18], que consiste en la adquisición de conjuntos de imágenes de entrenamiento correspondientes a los rostros de distintas personas. Con estas imágenes, Turk y Pentland proponen encontrar los componentes principales de la distribución de los rostros, o los vectores propios de la matriz de covarianza de cada uno de los conjuntos de imágenes de entrenamiento, tratando a cada imagen como un punto o vector en un espacio dimensional muy grande. Los vectores propios son ordenados, cada uno contribuyendo en diferente medida a la variación entre las imágenes de los rostros. Estos vectores pueden ser pensados como un conjunto de características que juntas engloban la variación entre imágenes. Cada ubicación en la imagen contribuye en mayor o menor medida a cada vector propio, que puede ser desplegado como una imagen "fantasmal" (etiquetada por Turk y Pentland en el caso de rostros como *eigenfaces*). Los n vectores propios con los mayores valores propios asociados contribuyen a la mayor varianza dentro del conjunto de imágenes y generan un espacio n-dimensional correspondiente a todas las posibles imágenes de un rostro de entrenamiento dado. El sistema puede ser utilizado para el reconocimiento de patrones al proyectar una nueva imagen a este espacio; si la nueva imagen corresponde al rostro utilizado para la definición del espacio, la proyección y la imagen serán muy parecidas entre sí, o en términos geométricos, su distancia euclidiana será lo suficientemente baja, y puede ser calculada, mediante la ecuación (1) como:

$$\varepsilon^2 = \theta - {\theta_f}^2 \ , \tag{1}$$

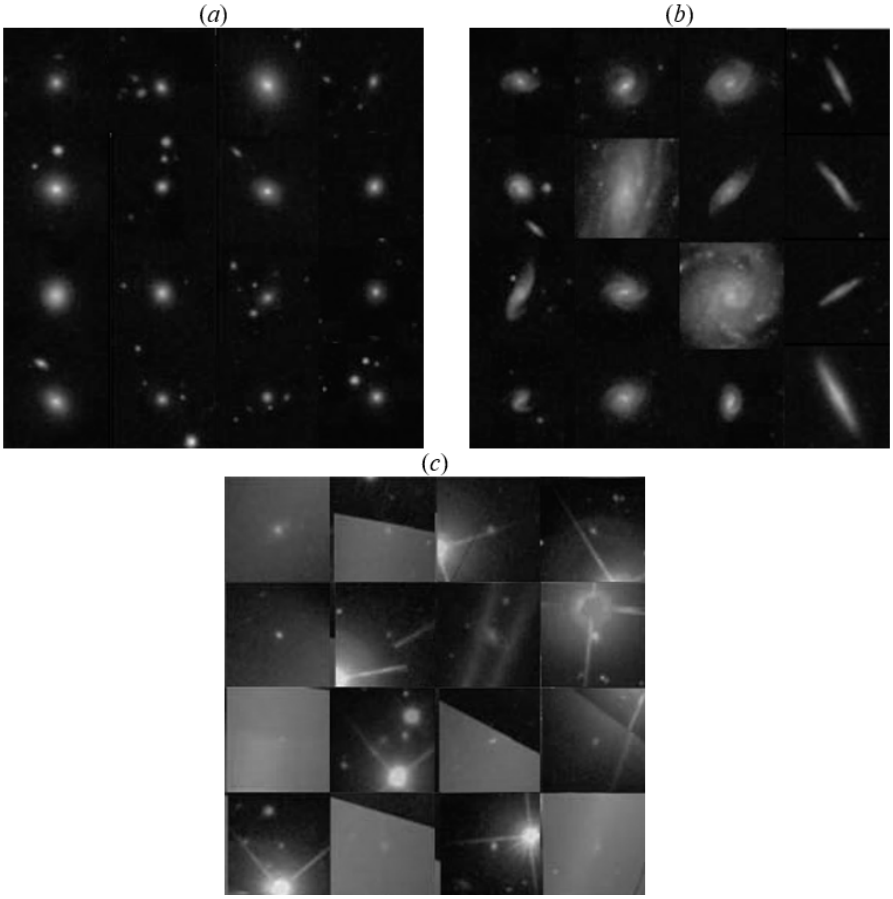

Fig. 1. Subconjuntos de 16 imágenes de cada tipo de galaxias (a) elípticas, (b) espirales y (c) inciertas, del conjunto total descargado de 504 imágenes.

donde ε representa la distancia, θ la imagen analizada (ajustada mediante la resta de la imagen promedio de las imágenes de entrenamiento), y θ_f el espacio definido por un conjunto de imágenes de un rostro determinado. Las 16 componentes principales asociadas con los 16 mayores valores propios para cada conjunto de galaxias se muestran en la Fig. 2. La mayor aportación a cada conjunto está dada por la imagen superior izquierda de cada conjunto, y disminuye de izquierda-derecha y arriba-abajo.

La clasificación de galaxias puede entenderse como un proceso similar al reconocimiento de rostros en el sentido de que ambos tipos de tareas implican procesamiento de alto nivel, para la cual la clasificación acorde a geometría detallada o información específica puede ser muy difícil, ineficiente, si no es que inútil en algunos casos. Con la intención de potenciar la eficiencia de ambos

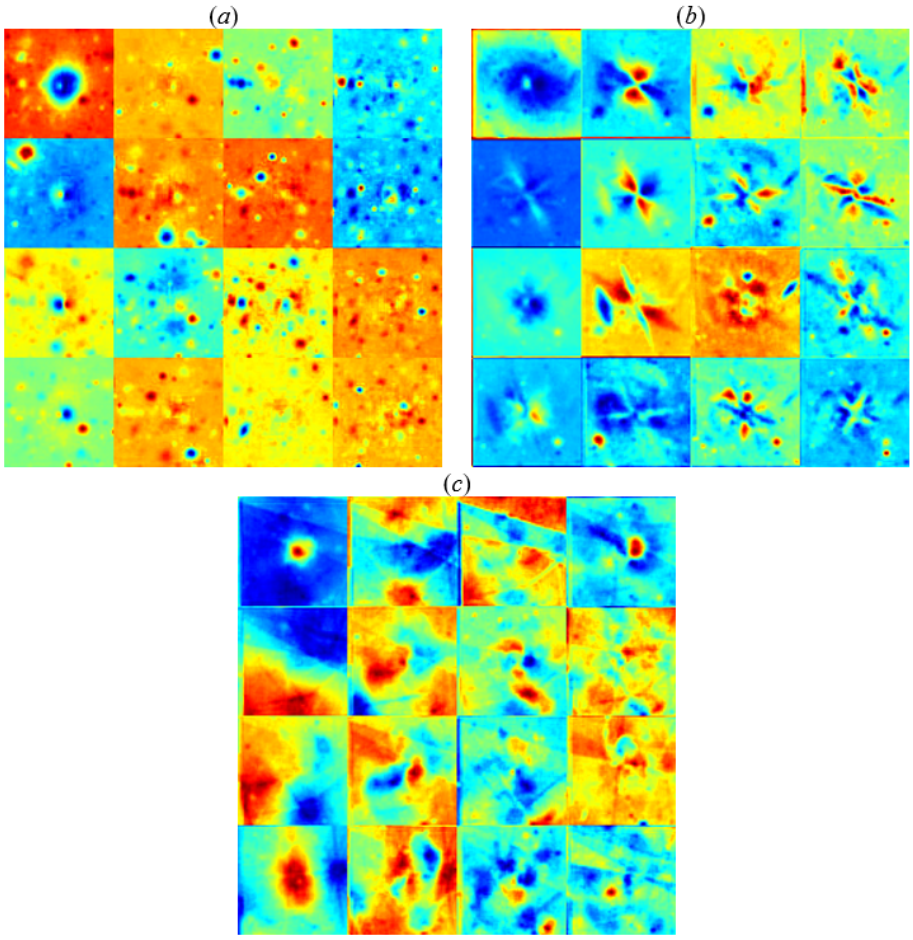

Fig. 2. Las 16 componentes principales para cada una de las tres clasificaciones propuestas, (a) elíptica, (b) espiral e (c) inciertas. En cada conjunto de imágenes, el componente principal es aquel en la esquina superior izquierda, y el orden de aportación disminuye de izquierda-derecha y arriba-abajo.

métodos de reconocimiento de patrones, RNA y ACP, en este trabajo se propone un algoritmo híbrido que potencia la eficiencia de la RNA preprocesando las imagenes de entrenamiento de la misma mediante ACP. El procesamiento consiste en enfatizar las características que hacen de una imagen correspondiente a una galaxia pertenecer a uno de los tres conjuntos: espirales, elípticas o inciertas. Esto se realiza aprovechando el hecho de que los vectores propios asociados con los mayores valores propios de cada conjunto de imágenes representan de manera muy general el patrón observado en el conjunto. Este patrón puede ser

incorporado a una imagen mediante la operación que añada a la imagen una combinación lineal ponderada de las componentes principales (ecuación (2)):

$$I_{\mathrm{N}} = \alpha I_{\mathrm{O}} + \frac{1}{2}(1-\alpha)\sum_{i=1}^{3} C_{i1}\left(1 - \frac{\varepsilon_i}{\sum_{j=1}^{3}\varepsilon_j}\right), \tag{2}$$

donde I_{N} representa la imagen modificada, I_{O} la imagen original, C_{i1} la primera componente principal (Fig. 2) o primer vector propio (asociado al mayor valor propio) del conjunto i de imágenes de galaxias, ε_i la distancia entre la imagen I_{O} y el espacio θ_i (Ecuación (1)), y α un parámetro entre 0 y 1. Un valor de $\alpha = 1$ no modificaría la imagen original, mientras que un valor de 0 completamente la reemplazaría por una combinación ponderada de las componentes principales. Las imágenes procesadas correspondientes a los subconjuntos de imágenes mostrados en la Fig. 1, se muestran en la Fig. 3.

2.3. Redes neuronales artificiales

Las Redes Neuronales Convolucionales (RNCs) o *convnets* [12] son una subclase de las RNAs con patrones de conectividad con restricciones entre algunas de las capas. Las RNCs pueden ser utilizadas cuando los datos de entrada exhiben algún tipo de estructura topológica [4], como el ordenamiento de pixeles en una malla o la estructura temporal de una señal de audio. Las RNCs contienen dos tipos de capas con conectividad restringida: capas convolucionales (*convolutional layers*) y capas de agrupamiento (*pooling layers*). Una capa convolucional toma una pila de mapas de características como una entrada, y convoluciona cada una de éstas con un conjunto de filtros para producir una pila de mapas de características de salida. Las RNCs típicamente tienen menos parámetros que las capas densas (o completamente conectadas) de otros tipos de RNAs. Debido a que las capas convolucionales son únicamente capaces de modelar correlaciones locales en la entrada, la dimensionalidad de los mapas de características es a menudo reducida entre capas convolucionales insertando capas de agrupamiento. Esto permite a las capas más altas modelar correlaciones a través de una parte más grande de la entrada. Al alternar capas convolucionales y de agrupamiento, las capas más altas en la red ven una representación progresivamente más burda de la entrada. De esta manera, estas capas son capaces de modelar abstracciones de mayor nivel más fácilmente porque cada unidad es capaz de "ver" una mayor parte de la entrada.

Se utilizó una RNC con quince capas, las cuales están distribuidas de la siguiente manera. La primera capa, la capa de entrada, es donde se especifica el tamaño de la imagen, que en este caso es de 69 por 69 por 1. Estos números corresponden a la altura, al ancho y al tamaño del canal. Los datos consisten en imágenes en escala de grises, por lo que el tamaño del canal es 1. La segunda capa es convolucional (misma que se repite en capas posteriores), y especifica el alto y el ancho de los filtros que usa la función de entrenamiento mientras escanea a lo largo de las imágenes, en este caso un filtro de 3 por 3. Esta capa siempre es seguida por la capa de normalización por lotes, la cual se encarga

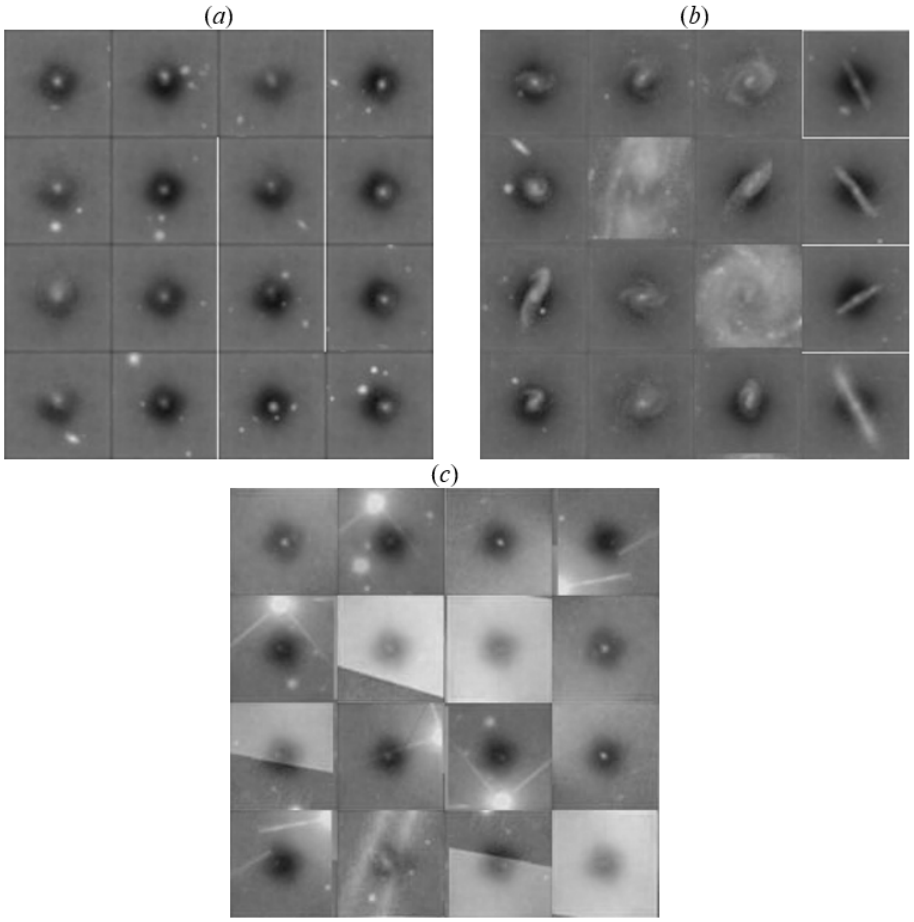

Fig. 3. Las imágenes preprocesadas de los subconjuntos de 16 imágenes de cada tipo de galaxias (a) elípticas, (b) espirales e (c) inciertas, mostradas en la Fig. 1.

de la normalización de las activaciones y de los gradientes que se propagan a través de la red. La capa de normalización por lotes va seguida de una función de activación no lineal. De igual manera se utilizaron dos capas de agrupación máxima, las cuales eliminan la información espacial redundante.

Se utiliza también una capa completamente conectada para combinar todas las características aprendidas por las capas anteriores en la imagen para identificar los patrones más grandes. Esta capa es seguida de la capa de Softmax, la cual normaliza la salida de la capa completamente conectada. Para finalizar se utilizó la capa de clasificación. Esta capa usa las probabilidades devueltas por la función de activación de Softmax para cada entrada, para asignar la entrada a una de las clases mutuamente excluyentes. El rendimiento promedio durante

100 corridas distintas de 200 épocas de entrenamiento cada una se muestra en la Fig. 4.

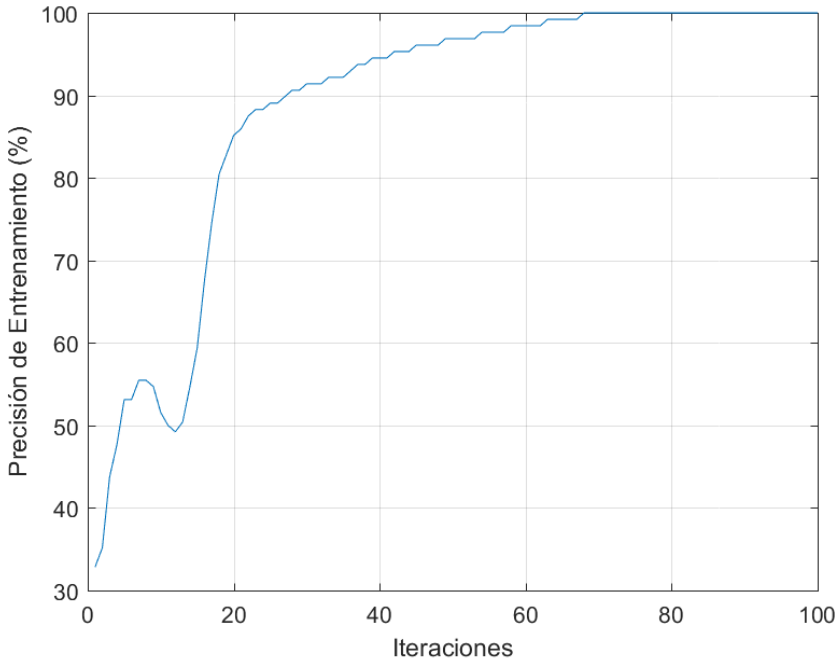

Fig. 4. El rendimiento promedio de la RNC considerando 100 corridas distintas de 200 épocas de entrenamiento cada una.

3. Conclusiones

En este trabajo se presenta el diseño y los resultados preliminares de un sistema basado en ACP y una RNA, particularmente una RNC para la clasificación de imágenes correspondientes a galaxias, con base en su morfología. Se obtuvieron imágenes de la SDSS, mismas que se preprocesaron utilizando un enfoque basado en ACP que enfatiza la categoría a la que pertenece cada imagen.

Estas imágenes se utilizaron para entrenar a una RNC utilizando las clasificaciones realizadas por un grupo masivo de voluntarios participantes en el proyecto Galaxy Zoo. La red es capaz de clasificar imágenes de galaxias directamente de los valores crudos de los pixeles, sin la necesidad de realizar extracción de características de forma manual.

4. Trabajo a futuro

A partir de estos resultados, se enfocarán esfuerzos en ampliar la cantidad de imágenes en la base de datos (de 504 a miles o millones de ellas), optimizando el preprocesamiento de las imágenes mediante ACP, y modificar la arquitectura de la RNC para mejorar el porcentaje de clasificación. De la misma manera, se propone realizar una serie de estudios estadísticos para validar la eficiencia del uso de las imágenes preprocesadas con ACP con respecto al uso de las imágenes sin procesar. Se pretende también robustecer la estructura del sistema de manera que permita la incorporación directa de nuevas y más extensas colecciones de imágenes disponibles al público.

Agradecimientos. Se agradece al Tecnológico Nacional de México/I.T. Mérida por el apoyo económico mediante los proyectos 6513.18-P y 6511.18-P. Los autores también agradecen el apoyo económico parcial de los proyectos 20181139, 20180472, 20181441, 20181028 y 20181141, así como al EDI, todos provistos por SIP/IPN.

Referencias

1. Galaxy zoo for astronomers homepage. `https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/about/faq`, last accessed 2018/04/15
2. Clery, D.: Galaxy Zoo volunteers share pain and glory of research. Science 333(6039), 173–175 (2011)
3. Conselice, C.J.: The relationship between stellar light distributions of galaxies and their formation histories. The Astrophysical Journal Supplement Series 147(1), 1 (2003)
4. Dieleman, S., Willett, K.W., Dambre, J.: Rotation-invariant convolutional neural networks for galaxy morphology prediction. Monthly notices of the royal astronomical society 450(2), 1441–1459 (2015)
5. Fadely, R., Hogg, D.W., Willman, B.: Star-galaxy classification in multi-band optical imaging. The Astrophysical Journal 760(1), 15 (2012)
6. GalaxyZoo Project: GalaxyZoo Project Data. `http://data.galaxyzoo.org`, last accessed 2018/03/15
7. Galton, F.: Finger Prints. Macmillan, London, 1st edn. (1892)
8. Hotelling, H.: Analysis of a complex of statistical variables into principal components. Journal of educational psychology 24(6), 417 (1933)
9. Hubble, E.P.: The realm of the nebulae, vol. 25. Yale University Press, New Haven, CO, 1st edn. (1936)
10. Kim, E.J., Brunner, R.J.: Star-galaxy classification using deep convolutional neural networks. Monthly Notices of the Royal Astronomical Society 464(4), stw2672 (2016)
11. Kim, E.J., Brunner, R.J., Carrasco Kind, M.: A hybrid ensemble learning approach to star–galaxy classification. Monthly Notices of the Royal Astronomical Society 453(1), 507–521 (2015)
12. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521(7553), 436 (2015)

13. Lintott, C.J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M.J., Nichol, R.C., Szalay, A., Andreescu, D., Murray, P., Vandenberg, J.: Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. Monthly Notices of the Royal Astronomical Society 389(3), 1179–1189 (2008)
14. Lisker, T.: Is the gini coefficient a stable measure of galaxy structure? The Astrophysical Journal Supplement Series 179(2), 319 (2008)
15. Nair, P.B., Abraham, R.G.: A catalog of detailed visual morphological classifications for 14,034 galaxies in the sloan digital sky survey. The Astrophysical Journal Supplement Series 186(2), 427 (2010)
16. Odewahn, S., Stockwell, E., Pennington, R., Humphreys, R., Zumach, W.: Automated star/galaxy discrimination with neural networks. In: Digitised Optical Sky Surveys, pp. 215–224. Springer, New York, NY (1992)
17. Orozco-Del-Castillo, M.G., Ortiz-Aleman, C., Martin, R., Avila-Carrera, R., Rodriguez-Castellanos, A.: Seismic data interpretation using the Hough transform and principal component analysis. Journal of Geophysics and Engineering 8(1), 61 (2010)
18. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of cognitive neuroscience 3(1), 71–86 (1991)
19. Weir, N., Fayyad, U.M., Djorgovski, S.: Automated star/galaxy classification for digitized POSS-II. The Astronomical Journal 109(6), 2401 (1995)