Galaxies image classification using artificial bee colony based on orthogonal Gegenbauer moments

Article *in* Soft Computing · September 2018

DOI: 10.1007/s00500-018-3521-2

Mohamed Elsayed abd el aziz
Zagazig University

Khalid Hosny
Zagazig University

Ibrahim M Selim
National Research Institute of Astronomy and Geophysics

Soft Computing
https://doi.org/10.1007/s00500-018-3521-2

METHODOLOGIES AND APPLICATION

Galaxies image classification using artificial bee colony based on orthogonal Gegenbauer moments

Mohamed Abd Elaziz[1] · **Khalid M. Hosny**[2] · **I. M. Selim**[3]

Abstract

The development of automated morphological classification schemes for galaxies is important to study the formation and subsequent evolution of galaxies in our universe. This paper proposed a new machine learning method for classifying three types of galaxies image (Hubble types) namely elliptical, lenticulars, and spirals. The proposed approach consists of three stages: In the first stage, the features are extracted from the galaxies image by using Gegenbauer moments that have several properties like scale, rotation, and invariant. However, not all these extract features are relevant; therefore, in the second stage, a swarm algorithm called artificial bee colony (ABC) is used as feature selection method, where ABC has a small number of parameters and its fast convergence to the global solution. The third stage is used to evaluate the performance of the selected features in the classification of galaxies image through using the support vector machine as a classifier. The experimental results are performed based on a sample from EFIGI catalog, and the results show the high performance of the proposed method.

Keywords Galaxy classification · Artificial bee colony (ABC) · Feature selection (FS) · Feature extraction · Orthogonal Gegenbauer moments · Support vector machine (SVM)

1 Introduction

At the beginning of the twentieth century, the scientists discover galaxies, what we now call spiral or elliptical galaxies, etc., were referred to as “spiral nebulae” and most astronomers believed them to be clouds of gas and stars associated with our own Milky Way (Sparke and Gallagher 2007). But in 1924 when Edwin Hubble was able to measure the distance to the “Great Nebula in Andromeda” (M31) and found its distance to be much larger than the diameter of the Milky Way, this meant that M31 is not the part from our Milky Way, but by extension it is other spiral nebulae, comparable to or even larger than the Milky Way. The space between galaxies is filled with a gas having an average density of less than one atom per cubic meter. One of the great discoveries of the twentieth century is that the universe is not static, but expanding; the galaxies all recede from each other, and from us (Sparke and Gallagher 2007).

So, we can say galaxies are the places where gas turns into luminous stars, gravitationally bound celestial entities composed of gas, dust, and billions of stars powered by nuclear reactions that also produce most of the chemical elements. Galaxies are the basic constituents of the universe at large scales and the building blocks of the large-scale structure of the universe. Approximately 200 billion or, more recently, at least 2 trillion galaxies exist in the observable universe, this is not the end of universe, but there are a lot of galaxies not discovered up to now. The majority of galaxies are gravitationally organized into associations known as galaxy groups (Selim et al. 2014). Understanding the galaxy formation and evolution is one of the main goals of modern cosmology. The ages, chemical composition and motions of the stars, also, the shapes that they make up, all of these properties give us more information about each galaxy’s past life. The scientists can discover more information about the composition and evolution of galaxies by using their morphology such as shape

Communicated by V. Loia.

✉ Mohamed Abd Elaziz
abd_el_aziz_m@yahoo.com

1 Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt

2 Department of Information Technology, Faculty of Computers and Informatics, Zagazig University, Zagazig 44519, Egypt

3 Astronomy Department, National Research Institute of Astronomy and Geophysics, Cairo, Egypt

Published online: 18 September 2018

and color. This information is used to evaluate the existing theories, also, to propose a new conjecture that can be used to discuss the physical operations governing star information, galaxies, and other information about the nature of the universe. However, the astrophysicists classify the galaxy to take the large benefit from this information since there exist a large number of catalogs. But this classification is performed manually using the visual inspection of photographs and this method has several drawbacks such as very slow and depends on the experience of astronomer that can lead to some misclassifications and inaccuracies (Hubble 1936; Ellison et al. 2008, 2011, 2013; Willett et al. 2017).

The process of determining the patterns and the trends of galaxies is considered the first stage in the empirical science. According to this process, Hubble presented a scheme to classify the galaxies. Recently, the galaxy families are defined using their physical characteristics and basic correlations that explain their formative and physics histories Galaxy Zoo. However, the morphological analysis of galaxies has more attraction since it can be used to understand the computational processes inside these objects based on the precious clues (Willett et al. 2017). As well as, the Sloan Digital Sky Survey (SDSS) program provides the astronomy by large amount of data, where the SDSS has images of a large portion of the sky and found more than 800 million galaxies. Classifying them by eye would take huge amount of time. SDSS cleverly called upon the public to help and asked volunteers to look at images of new galaxies. In order to avoid these problems, several machine learning (ML) methods have been proposed to deal with astronomical image. For example, artificial neural networks (ANNs) are automated classification of SDSS of data (Bazarghan and Gupta 2008; Kheirdastan and Bazarghan 2016). Also, the convolutional neural network is used to reconstruct astronomical image (Flamary 2016) and an efficient and robust NN training tool for machine learning in astronomy (Graff et al. 2014). Also, the ANNs have been applied to classify the galaxy images as in Storrie-Lombardi et al. (1992) and the ANN can also classify galaxy based on stellar spectra as in Von Hippel (1994), Bailer-Jones et al. (1998), Ball et al. (2004), Almeida et al. (2010). Additionally, Dieleman et al. (2015) used the deep neural network (DNN) for Galaxy Zoo classification in which the main objective of this method is to minimize the sensitivity to changes in the translation, rotation, and scaling. The accuracy of DNN method is 99% with respect to the classification performed by human; however, this accuracy is not perfect since several errors can be combined with human classification, so the same errors are found in DNN (Ferrari et al. 2015). The previous neural networks have been used for features extracted from the images as well as raw pixel data as inputs to the neural networks. This will not only decrease classification error, but also allow astronomers to pursue more stimulating tasks.

In Gauci et al. (2010), the authors introduced a machine learning method for morphology classification of two types of galaxy images, spiral galaxies and elliptical galaxies. This method is based on the decision tree learning algorithms such as CART, the C4.5, the random forest, and fuzzy inference systems. Also, Freeman et al. (2013) used the random forest method for classification of 1639 *HST*/WFC3 galaxies image, which used the intensity, multimode, and deviation statistics to identify the disturbed morphologies. Polsterer et al. (2015) present an unsupervised approach to classify a set of galaxy images through generating a set of features. This method is called parallelized rotation/flipping invariant Kohonen maps (PINK), and it is used a multi-core CPU/GPU environment. Also, this method is based on an improvement of self-organizing maps with intensive rotation and flipping invariant similarity measure.

In Huertas-Company et al. (2008, 2009), the classification is performed using the support vector machines, in which this method is used for high-redshift galaxies. This method has several advantages to deal with large number of parameters; also, it assigns probabilities rather than binary classes as in Huertas-Company et al. (2010). The method in Huertas-Company et al. (2010) is used to revisit the Hubble sequence in the SDSS DR7 spectroscopic sample that contains four types (E, S0, Sab, and Scd), and this method extracts the features like shape, concentration, and color from the galaxy images; then, the support vector machine is used to perform the classification. Abraham et al. (2003) used the Gini index as a measure to find the morphology, and there are several other methods for galaxy morphologies such as GALFIT (Peng et al. 2002), Ganalyzer (Shamir 2011; Schutter and Shamir 2015), and SpArcFiRe (Davis and Hayes 2014).

Moreover, the MORFOMETRYKA method is proposed as galaxy classification using the linear discriminant analysis algorithm (Ferrari et al. 2015). This method extracted different features such as asymmetry, concentration, spirality, smoothness, and entropy from elliptical and spiral images and then classifies them which gives results nearly 90%. In other words, many machine learning approaches have been used to extract a photometric data from galaxy image and then used these features for classification. For example, in Cavuoti et al. (2013), there are four machine learning methods, namely support vector machines, multi-layer perceptron based on the scaled conjugate gradient, conjugate gradient, and the quasi-Newton learning. Also, the quasi-Newton algorithm is used to estimate the photometric redshifts (Cavuoti et al. 2012).

In addition, the nonnegative matrix factorization is, recently, used as feature extraction method to improve the classification of galaxy images as in Selim and Abd El Aziz (2017) and after that the MAX rule is used to predict the label of testing images. The accuracy of this method is reached nearly ~ 92% for sample containing 700 images. However, this method changes the nature of the data, as well as, they not captural the high information from the image. The modified version of sine–cosine algorithm (SCA) is used by Abd El Aziz et al. (2017) to improve the classification of 5000 images, in which the SCA is used as feature selection method to select the relevant features from the set of extracted features using shape, texture, and color extracted methods.

In this paper, an alternative method to classify galaxy images is proposed. In this method, the orthogonal Gegenbauer moments (Hosny 2011) are used to extract the features and the artificial bee colony (Karaboga and Akay 2009; Karaboga and Ozturk 2011) is utilized as a feature selection method.

The Gegenbauer moments are orthogonal moments defined in Cartesian coordinates over the domain $[-1, 1] \times [-1, 1]$ (Hosny 2014). These moments have attractive properties such as (1) representation of digital images with minimum information redundancy. (2) Capture of image features due to the distribution of roots over the entire domain. (3) Robustness against different kinds of noise. During the last decade, Gegenbauer moments are used in several applications. Liao and Chen (2013) used low-order Gegenbauer moments in object recognition. Wang and Mottershead (2013) used the moment descriptors in full-field strain and displacement measurements. Habibzadeh et al. (2014) used Gegenbauer features in white blood cell differential counts in low-resolution images. Salouan et al. (2015) used Gegenbauer moments in recognition of handwritten Arabic characters. Wang et al. (2016) presented a short review paper showing the importance of Gegenbauer moment invariants in different applications.

In order to ignore irrelevant features, a feature selection process is performed. The artificial bee colony (ABC) is used in this process (Hancer et al. 2015). The ABC is a swarm algorithm which simulates the behavior of bee colony. The ABC has few parameters; therefore, it has been used in several applications such as clustering (Ji et al. 2015), data mining (Çelik Mete et al. 2011), and image segmentation (Ma et al. 2011). The population of ABC contains three classes of bee, namely employed, onlooker, and scouts. The bees in first class (employed bees) are going to the food source (solution) and transfer the information that obtained about the nectar amount of a food source (fitness function value) to the second class of bees which called onlooker bees. The responsibility to choose the best solution from the first class of bees (employed) is performed by those onlooker bees. The third class of bees (scout bees) is responsible to determine a new solution (in random form) in the case there is no improvement in the current solution.

In this paper, we proposed a new approach to classification of galaxy images. The proposed approach consists of three stages: In the first stage, the Gegenbauer moments are used to extract the discriminate features from the image; then, in the second stage the artificial bee colony (ABC) is used to select the most important features from the first stage. The last stage is the classification stage in which the support vector machine (SVM) is used as classifier algorithm.

The rest of the paper is organized as follows: Sect. 2 introduces the basic concepts of the orthogonal Gegenbauer moments. In Sect. 3, the artificial bee colony algorithm is introduced. In Sect. 4, the proposed method for classification of galaxy image is illustrated. In Sect. 5, the results of the proposed algorithm are discussed. The conclusion and the future works are given in Sect. 6.

2 Orthogonal Gegenbauer moments

Gegenbauer moments are orthogonal moments defined in Cartesian coordinates over the domain $[-1, 1] \times [-1, 1]$:

$$A_{n,m} = \frac{1}{C_n(\alpha)C_m(\alpha)} \int_{-1}^{1}\int_{-1}^{1} f(x, y)G_n^{(\alpha)}(x)G_m^{(\alpha)}(y)w^{(\alpha)} \times (x)w^{(\alpha)}(y)\mathrm{d}x\,\mathrm{d}y \tag{1}$$

where $n, m = 0, 1, 2, 3, \ldots$ and $G_n^{(\alpha)}$ is the orthogonal Gegenbauer function with the scaling parameter; $\alpha > -0.5$ is defined as (Abramowitz et al. 1965):

$$G_n^{(\alpha)}(x) = \sum_{k=0}^{\lfloor \frac{n}{2} \rfloor} B_{n,k}^{(\alpha)} x^{n-2k} \tag{2}$$

with the coefficient matrix:

$$B_{n,k}^{(\alpha)} = (-1)^k \frac{\Gamma(n-k+\alpha)2^{n-2k}}{k!\,(n-2k)!\,\Gamma(\alpha)} \tag{3}$$

where $\frac{n}{2}$ and $\Gamma(\cdot)$ refer to the floor operator and the gamma function, respectively. The floor function, $\frac{\boldsymbol{n}}{2}$, is equal to $\frac{(\boldsymbol{n}-1)}{2}$ for odd values of $\boldsymbol{n}$ and/or equal to $\frac{\boldsymbol{n}}{2}$ for even values of $\boldsymbol{n}$.

The orthogonality property is:

$$\int_{-1}^{1} G_n^{(\alpha)}(x)G_m^{(\alpha)}(x)w^{(\alpha)}(x)\mathrm{d}x = C_n(\alpha)\delta_{nm} \tag{4}$$

where δ_{nm} is the Kronecker function; $w^{(\alpha)}$ is the weight function and defined as:

$$w^{(\alpha)}(x) = \left(1 - x^2\right)^{\alpha-0.5} \tag{5}$$

The normalization constant $C_n(\alpha)$ is defined as:

$$C_n(\alpha) = \frac{2\pi\Gamma(n+2\alpha)}{2^{2\alpha}n!\,(n+\alpha)[\Gamma(\alpha)]^2} \tag{6}$$

For a maximum order max, the number of independent Gegenbauer moments is:

$$N\text{max} = \frac{(\text{max}+1)(\text{max}+2)}{2} \tag{7}$$

2.1 Exact Gegenbauer moments

Based on the method proposed in Hosny (2011), highly accurate Gegenbauer moments are obtained using the following form:

$$A_{n,m} = \frac{1}{C_n(\alpha)C_m(\alpha)} \sum_{i=1}^{M}\sum_{j=1}^{N} T_{nm}\left(x_i, y_j\right) f\left(x_i, y_j\right) \tag{8}$$

where

$$T_{nm}\left(x_i, y_j\right) = \int_{x_i-\frac{\Delta x_i}{2}}^{x_i+\frac{\Delta x_i}{2}} \int_{y_j-\frac{\Delta y_j}{2}}^{y_j+\frac{\Delta y_j}{2}} G_n^{(\alpha)}(x) G_m^{(\alpha)}(y) w^{(\alpha)} \times (x) w^{(\alpha)}(y) \mathrm{d}x\mathrm{d}y \tag{9}$$

Double integration represented by Eq. (9) could be evaluated separately. For a digital image of size $M \times N$, the discrete set of points $\left(x_i, y_j\right) \in [-1, 1] \times [-1, 1]$ are defined as follows:

$$x_i = -1 + \left(i - \frac{1}{2}\right)\Delta x; \quad y_j = -1 + \left(j - \frac{1}{2}\right)\Delta y \tag{10}$$

where $i = 1, 2, 3, \ldots, M$, $j = 1, 2, 3, \ldots, N$, $\Delta x_i = 2/M$, and $\Delta y_j = 2/N$.

Substituting Eq. (9) in (8) yields:

$$A_{n,m} = \frac{1}{C_n(\alpha)C_m(\alpha)} \sum_{i=1}^{M}\sum_{j=1}^{N} IX_n(x_i) IY_m\left(y_j\right) f\left(x_i, y_j\right) \tag{11}$$

where the kernels are defined as follows:

$$IX_n(x_i) = \int_{x_i-\frac{\Delta x_i}{2}}^{x_i+\frac{\Delta x_i}{2}} G_n^{(\alpha)}(x) w^{(\alpha)}(x) \mathrm{d}x \tag{12}$$

$$IY_m\left(y_j\right) = \int_{y_j-\frac{\Delta y_j}{2}}^{y_j+\frac{\Delta y_j}{2}} G_m^{(\alpha)}(y) w^{(\alpha)}(y) \mathrm{d}y \tag{13}$$

Substituting $G_n^{(\alpha)}(x)$ and $w^{(\alpha)}(x)$ in Eq. (12) yields:

$$IX_n(x_i) = \sum_{k=0}^{\lfloor \frac{n}{2} \rfloor} B_{n,k}^{(\alpha)} \int_{x_i-\frac{\Delta x_i}{2}}^{x_i+\frac{\Delta x_i}{2}} x^{n-2k}\left(1-x^2\right)^{\alpha-0.5} \mathrm{d}x \tag{14}$$

Equation (14) is rewritten in a compact form as follows:

$$IX_n(x_i) = \sum_{k=0}^{\lfloor \frac{n}{2} \rfloor} B_{n,k}^{(\alpha)} I_{n-2k}(x_i) \tag{15}$$

Similarly,

$$IY_m\left(y_j\right) = \sum_{t=0}^{\frac{m}{2}} B_{m,t}^{(\alpha)} I_{m-2t}\left(y_j\right) \tag{16}$$

Analytical evaluation of the definite integrations $IX_n(x_i)$ and $IY_m\left(y_j\right)$ results in exact computation of the kernels, $IX_n(x_i)$ and $IX_m\left(y_j\right)$.

2.2 Highly accurate Gegenbauer moment invariants

Invariance to the similarity transformations, rotation, scaling, and translation (RST) is very essential property in pattern recognition applications. Gegenbauer moment invariants (GMIs) with respect to RST could be computed through direct method where Gegenbauer polynomials are expressed in terms of RST geometric/regular moment invariants. GMIs of order $(n+m)$ are defined as follows (Hosny 2014):

$$\mathrm{GMI}_{n,m} = \frac{1}{C_n(\alpha)C_m(\alpha)} \times \sum_{\substack{k=0 \\ n-k=\text{even}}}^{n} \sum_{\substack{\ell=0 \\ m-\ell=\text{even}}}^{m} B_{n,k}^{(\alpha)} B_{m,\ell}^{(\alpha)} \mathrm{RMI}_{(n-2k),(m-2\ell)} \tag{17}$$

where the regular moment invariants $\mathrm{RMI}_{p,q}$ for an image function $f(x, y)$ with an order $(p+q)$ are defined as follows:

$$\mathrm{RMI}_{pq} = \frac{1}{\mu_{00}^{\lambda}} \sum_{k=0}^{p}\sum_{m=0}^{q} \binom{p}{k}\binom{q}{m} (-1)^m \times (\sin\theta)^{k+m} (\cos\theta)^{p+q-k-m} \mu_{p-k+m,q-m+k} \tag{18}$$

With a rotation angle defined as:

$$\theta = \frac{1}{2}\tan^{-1}\left(\frac{2\mu_{11}}{\mu_{20}-\mu_{02}}\right) \tag{19}$$

The central moments are:

$$\mu_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - x_0)^p (y - y_0)^q h(x, y)\mathrm{d}x\mathrm{d}y \tag{20}$$

With:

$$x_0 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} x_i h(x_i, y_j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} h(x_i, y_j)},$$
$$y_0 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} y_j h(x_i, y_j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} h(x_i, y_j)} \tag{21}$$

$$h(x, y) = \left(1 - x^2\right)^{\alpha-0.5}\left(1 - y^2\right)^{\alpha-0.5} f(x, y) \tag{22}$$

$$\lambda = \frac{p + q + 2}{2} \tag{23}$$

The use of the GMIs requires conversion of the 2D matrix of Gegenbauer moments into a 1D vector. The conversion process could be done by using the conversion code (Hosny 2007).

3 Artificial bee colony (ABC)

In this section, the basic steps of the ABC algorithm are introduced based on the three class of the population. The algorithm begins by generating a random population of size N (each individual represents as solution $x_i \in R^d, i = 1, 2, \ldots, N$). The position of employed bee x_i is updated based on its neighbor bee x_k which selected random as in Eq. (24) (Karaboga and Akay 2009):

$$v_{ij} = x_{ij} + \varphi_{ij}\left(x_{ij} - x_{kj}\right), k = \mathrm{int}(\mathrm{rand} \times N),$$
$$j = 1, \ldots, d, \varphi_{ij} \in [-1, 1] \tag{24}$$

Then, the fitness functions $f(v_i)$ and $f(x_i)$ for both solutions v_i and x_i are computed, respectively, and if $f(v_i)$ is better than $f(x_i)$, then x_i is removed from memory; at the same time v_i is added to the memory; otherwise, x_i remains in the memory.

Thereafter, the probability of each bee is computed using Eq. (25) (Dieleman et al. 2015):

$$P_i = \frac{\mathrm{fit}_i}{\sum_{i=1}^{N} \mathrm{fit}_i}, \quad \mathrm{fit}_i = \begin{cases} \frac{1}{1+f_i} & f_i > 0 \\ 1 + \mathrm{abs}(f_i) & \text{otherwise} \end{cases} \tag{25}$$

Based on these probability values, the onlooker bees select the employed bee which has the higher P [this selection is performed by using the roulette wheel selection mechanism (Hancer et al. 2015)].

Then, using the same approach that the employed bees are applied to update their position, the onlookers update their position. The fitness function for each onlooker is computed and compared with the old fitness function value of employed bees to decide whether the old solution removed or remained in the memory.

The solutions are not improved after some iterations; then, they become abandoned solution and the third class of bees (scout looks) is used to replace them as in equation:

$$x_{ij} = x_j^{\min} + \left(x_j^{\max} - x_j^{\min}\right) * \mathrm{rand}, \tag{26}$$

where x_{ij} is a parameter to be optimized for the ith employed bee, $x_j^{\max}$ and $x_j^{\min}$ are the upper and lower bounds for x_{ij}, respectively, and rand is a random number. After generating a new solution x_{ij} then becomes employed bee.

4 The proposed prediction system

The proposed system for classification of the galaxy images is illustrated in Algorithm 1 which consists of three stages. In the first stage, the exact Gegenbauer moments are used to extract the discriminate features from each image I, then the matrix of all features M_{feat} (the size of M_{feat} is $Q \times D$ where Q is the number of images and D is the number of features) constructed. In the second stage, the ABC algorithm is used to select the best features from M_{feat}. This matrix is used as input to ABC, and the output is the best solution x_{best} which is represented by binary vector. The ABC is started by generating random population of size $N \times D$ (N is the bees number). The position x of each employed bee is updated based on Eq. (24); then, the solution is converted into a binary as follows:

```
Algorithm 1: Proposed System
Input: D : Galaxy images data set and y:decision feature
Output: Selected features, Performance of system
First Stage (features Extracting).
M_feat=Gegenbauer Moment (D)
Second Stage (select features based ABC).
 1. Initialize the values of maxiter: maximum number of iterations,
 2. Generate randomly a population of size K.
 3. t = 1;
 4. While(t ≤ t_max)
 5. For i = 1:N  //  Employee bees
 6.        Update the solution x_i(Eq. (24)).
 7.        Evaluate the Fitness function F_i (Eq. (28)).
 8. End For
 9. Find the best fitness function F_best and the best solution x_best
 10. For i = 1:N  // Onlooker bees
 11.       Update the solution x_i(Eq. (24)).
 12.       Evaluate the Fitness function F_i (Eq. (28)).
 13. End For
 14. Update F_best and x_best
 15. Determine the abandoned solutions
 16. For j=1:size(abandoned solutions) // Scout bee
 17.         Remove the abandoned solution and add new random solution (Eq. (26)).
 18. End For
 19. End While
Third Phase classification
 1. D_new = D(x_best) // choose the columns corresponding to 1's in x_best.
 2. Split D_new into training set and testing set.
 3. Train the SVM based on the training set.
 4. Apply the SVM trained model to the testing set and compute the output
 5. Evaluate the performance of proposed approach.
```

$$x(t) = \begin{cases} 1 \text{ if } x(t) \geq \sigma \\ 0 \text{ otherwise} \end{cases} \tag{27}$$

For each solution the fitness function F is computed as:

$$F_i = \frac{Q_C}{Q}, \quad i = 1, \ldots, N \tag{28}$$

where Q_C represents the correct number of classified images.

The best fitness F_{best} is determined with the best position x_{best}. Then, the position and the fitness function of both the onlooker and scout bees are updated and the best fitness function also updated. These steps are performed until the stopping conditions are satisfied. In the final stage, the SVM is used as classification algorithm in which the data are divided into training and testing set after updating the dataset based on the selected features. The training set is used in the learning phase, while the testing set is used to evaluate the learned model of SVM and compute the performance between the target and the predicted value.

Table 1 Features of ellipticals, lenticulars, and spirals galaxies

Galaxy type	Feature
Spiral	The spherical structure in the center of object is called bulge The younger stars, gas, and dust formed the disk that made up the arm structures (spiral) The spherical structure around the disk and bulge is called the halo which contains the old stars clusters
Elliptical	This type similar to spiral which has a central bulges and disks except no spiral arms
Lenticular	This type has completed 2D symmetry; also it extends along one axis longer than another axis

Table 2 Set of performance measures

Measure	Formula
Accuracy (Acc)	$\text{Acc} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \times 100$
Recall (Rec)	$\text{Rec} = \frac{\text{TP}}{\text{TP+FN}} \times 100$
Specificity (Spec)	$\text{Spec} = \frac{\text{TN}}{\text{FP+TN}} \times 100$
F measure	$F = \frac{2(\text{precision+recall})}{\text{precision} \times \text{recall}}$

5 Experimental results

5.1 Data description

The sample of dataset contains 711 galaxies taken EFIGI catalog to assess the accuracy of the proposed method (Baillard et al. 2011). The EFIGI is a large galaxy database which is presented by modern imaging surveys. These images of EFIGI are extracted from SDSS DR4 and PGC, and the EFIGI was used as a multi-wavelength reference database of the morphological characteristics of nearby galaxies. This catalog has many characteristics which combined images from several other catalogs such as NASA Extragalactic Database, Sloan Digital Sky Survey, Value-Added Galaxy Catalogue, HyperLeda, and Principal Galaxy Catalogue. In addition to all types of Hubble involved in the EFIGI catalog, it is a dense subsample of the local universe. For more information, see Peng et al. 2002; Davis and Hayes 2014.

From this catalog we select three types: spirals, ellipticals, and lenticulars and galaxies, and each type has number of images as 900, 118, and 29 images, respectively. The main properties of these three types are decrypted in Table 1.

Table 3 The performance results of the proposed method at different number of features

Classifier	Accuracy	*F* measure	Recall	Precision	Number of features
NB	90.42	65.97	68.06	64.82	20
SVM	90.80	60.42	63.41	59.36	
NB	91.19	76.89	93.34	68.18	70
SVM	91.98	65.85	72.54	61.61	
NB	92.34	76.89	93.34	68.18	150
SVM	94.63	87.21	96.32	81.30	
NB	89.27	66.87	70.33	64.38	332 (all features)
SVM	90.04	47.59	55.88	44.83	

5.2 Performance measures

In this section the set of performance measures which are used to assess the accuracy of proposed approach defined as given in Table 2, where FP, TP, FN, and TN are the false positive, true positive, false negative, and true negative, respectively. Also, the Naïve Bayes classifier is used also to evaluate the performance of the selected features and compare the results obtained by the SVM with it. The parameters of ABC are set as in the original reference (Hancer et al. 2015), the number of iterations is set to 500, and the algorithm is run 35 times. The experiments were implemented in MATLAB and run in windows environment with 64-bit support.

5.3 Results and discussion

The dataset is split randomly into two: training set (75%) and testing set (25%). The results of the proposed approach to classify the dataset are given in Table 3 and Figs. 1 and 2, in which there are three different sets of selected features 20, 70, and 150 in addition to the set of all features.

Figure 1 shows the visualization of the confusion matrices at the four sizes of features, in which this figure provides a fast comparison between the accuracy of the two classifiers that have different features. A confusion matrix is very important since the other measures are computed according to it. Therefore, from Fig. 1 we can be observed that when the of the proposed method selects only 20 features the SVM is better than the NB according to the accuracy; however, the *F* measure, recall, and precision of the SVM are less than those of the NB classifier. The same results are obtained in the case 70 features were selected, in which the SVM has accuracy better than NB, but the NB has better performance according to the other terms. Moreover, in the case of selected 150 features, the SVM has better results than the NB in terms of all measures; also, the performance of both classifiers is better than their performance in the other selected features (20, 70, and also all features). In addition, from this figure, we can see that the two classifiers have good performance, also, both of them can distinguish between the images of the first class and the other two classes. However, this performance is decreased in the case of the second class, since the NB and SVM cannot separate 15 and 12 images in their best case (when 70 features were selected). For the third class the SVM has higher performance than NB to separate the elements of this class from the other classes.

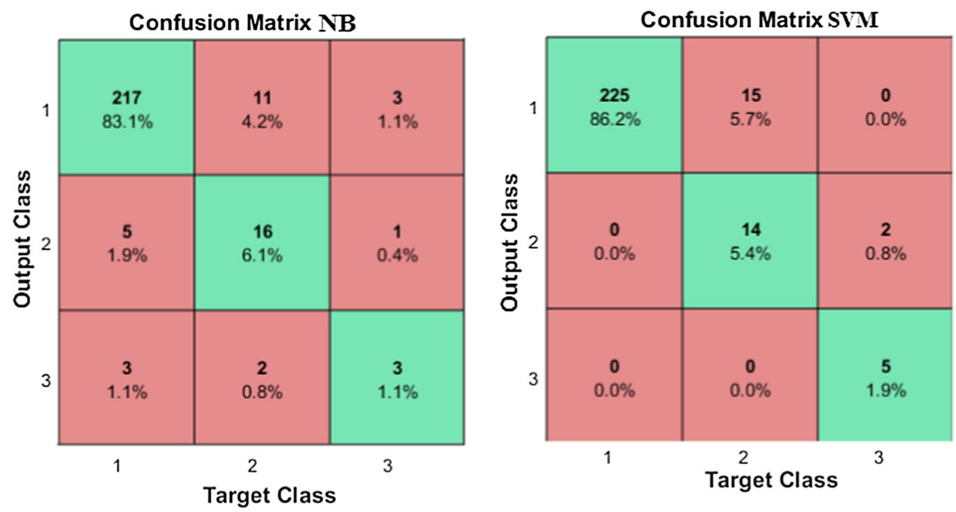
From Table 3, we can observe that the best accuracy is achieved by the number of 150 features nearly the half of set which is 92.65, 94.63 for 94.63 and SVM, respectively. These results are better than the other three cases (20, 70, and ALL); however, the set that contains only 70 features is better than the other two sets (its accuracy is 91.22 and 91.98 for NB and SVM, respectively). Also, the set of 20 features is better than the all features.

Moreover, Fig. 2 shows the average of accuracy over all the classifiers; from this figure, we can see that classification accuracy of the 150 features is the better which is consistent with Table 3. In addition, from Fig. 2b, we can conclude that the best classifier is the SVM which achieves the accuracy equal to 91.96.

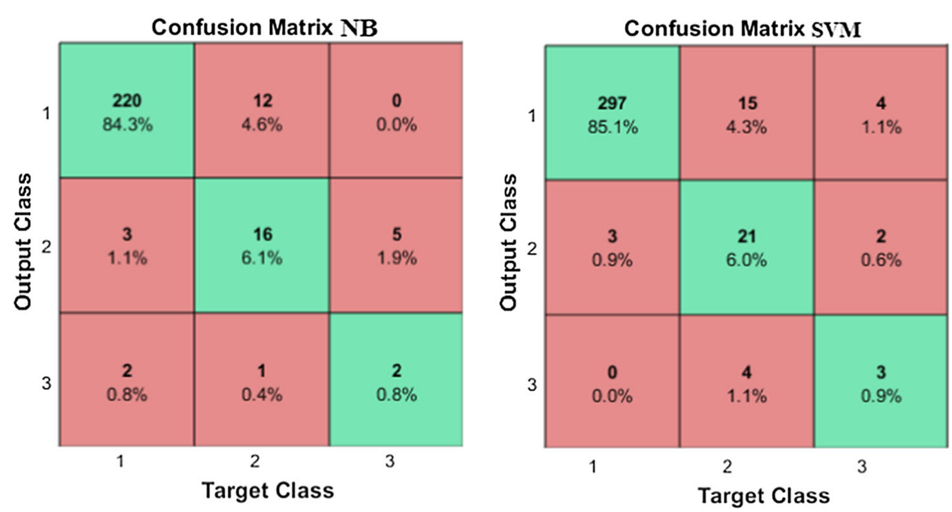
6 Conclusions

This paper provides a new classification system for galaxies images which consists of three stages each of them performing a specific task, the first stage extracts the features from the galaxies image by using Gegenbauer moments which has several properties like scale, rotation, and invariant. However, not all these extract features are relevant; therefore, in the second stage the artificial bee colony is used as feature selection method. The artificial bee colony has small number of parameters and its fast convergence to global solution. The third stage is used to evaluate the performance of the selected features in the classification of galaxies image.

The experimental results show that the proposed method has good performance results to classify the galaxy image.



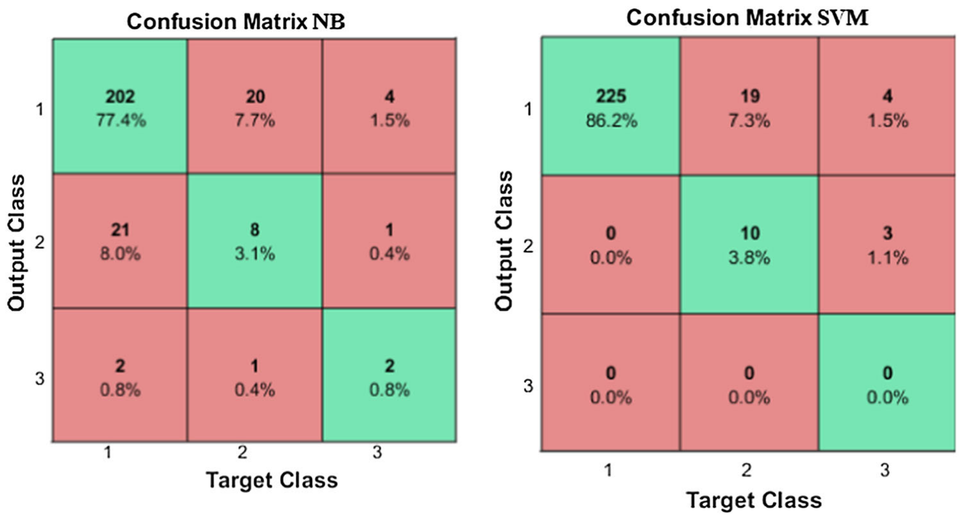
(A) At 20 selected features



(B) At 70 selected features

(C) At 150 selected features

Fig. 1 The confusion matrices at different selected features using two classifiers



(D) At ALL features

Fig. 1 continued

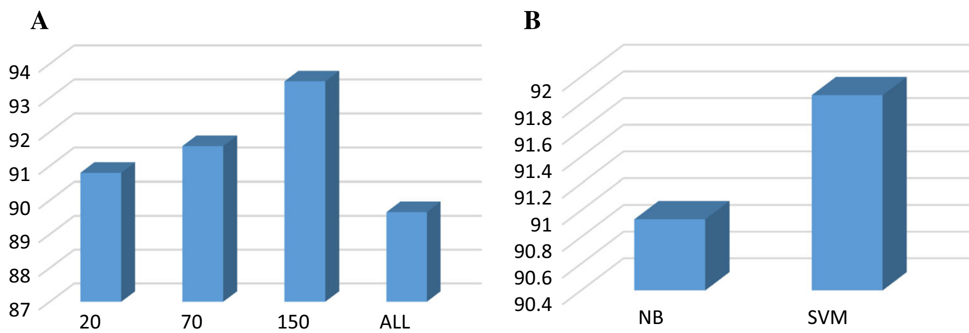


Fig. 2 The comparison between accuracy of **a** the proposed method along different number of features. **b** The two classifiers NB and SVM

The results achieve the best fit of adapted range of morphological classification with the smallest value of error and the comparisons show that the proposed algorithm works more efficient and faster.

In the future works the proposed system can be applied to a large dataset and can be enhanced by improving the performance of the ABC algorithm or used another feature extraction method.

Funding This research received no specific grant from any funding agency in public, commercial, or not-for-profit sectors.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

Abd El Aziz M, Selim IM, Xiong S (2017) Automatic detection of galaxy type from datasets of galaxies image based on image retrieval approach. Sci Rep. https://doi.org/10.1038/s41598-017-04605-9

Abraham RG, van den Bergh S, Nair P (2003) A new approach to galaxy morphology. I. Analysis of the Sloan Digital Sky Survey early data release. Astrophys J 588:218–229. https://doi.org/10.1086/373919

Abramowitz M, Stegun IA, Miller D (1965) Handbook of mathematical functions with formulas, graphs and mathematical tables (National Bureau of Standards Applied Mathematics Series No. 55). J Appl Mech 32:239

Almeida C, Baugh CM, Lacey CG et al (2010) Modelling the dusty universe—I. Introducing the artificial neural network and first applications to luminosity and colour distributions. Mon Not R Astron Soc 402:544–564. https://doi.org/10.1111/j.1365-2966.2009.15920.x

Bailer-Jones CAL, Irwin M, Von Hippel T (1998) Semi-automated extraction of digital objective prism spectra. Mon Not R Astron Soc 298:1061–1068. https://doi.org/10.1111/j.1365-8711.1998.01746.x

Baillard A, Bertin E, de Lapparent V et al (2011) The EFIGI catalogue of 4458 nearby galaxies with detailed morphology. Astron Astrophys. https://doi.org/10.1051/0004-6361/201016423

Ball NM, Loveday J, Fukugita M et al (2004) Galaxy types in the Sloan Digital Sky Survey using supervised artificial neural networks. Mon Not R Astron Soc 348:1038–1046. https://doi.org/10.1111/j.1365-2966.2004.07429.x

Bazarghan M, Gupta R (2008) Automated classification of Sloan Digital Sky Survey (SDSS) stellar spectra using artificial neural networks. Astrophys Space Sci 315:201–210. https://doi.org/10.1007/s10509-008-9816-5

Cavuoti S, Brescia M, Longo G, Mercurio A (2012) Photometric redshifts with the quasi Newton algorithm (MLPQNA) results in the PHAT1 contest. Astron Astrophys 546:A13. https://doi.org/10.1051/0004-6361/201219755

Cavuoti S, Brescia M, D'Abrusco R et al (2013) Photometric classification of emission line galaxies with machine-learning methods. Mon Not R Astron Soc 437:968–975. https://doi.org/10.1093/mnras/stt1961

Çelik M, Karaboğa D, Köylü F (2011) Artificial bee colony data miner (abc-miner). In: International symposium on innovations in intelligent systems and applications (INISTA), 2011. IEEE

Davis DR, Hayes WB (2014) SpArcFiRe: scalable automated detection of spiral galaxy arm segments. Astrophys J 790:1–4. https://doi.org/10.1088/0004-637X/790/2/87

Dieleman S, Willett KW, Dambre J (2015) Rotation-invariant convolutional neural networks for galaxy morphology prediction. Mon Not R Astron Soc 450:1441–1459. https://doi.org/10.1093/mnras/stv632

Ellison SL, Patton DR, Simard L, McConnachie AW (2008) Galaxy pairs in the Sloan Digital Sky Survey. I. Star formation, active galactic nucleus fraction, and the mass-metallicity relation. Astron J 135:1877–1899. https://doi.org/10.1088/0004-6256/135/5/1877

Ellison SL, Patton DR, Mendel JT, Scudder JM (2011) Galaxy pairs in the Sloan Digital Sky Survey—IV. Interactions trigger active galactic nuclei. Mon Not R Astron Soc 418:2043–2053. https://doi.org/10.1111/j.1365-2966.2011.19624.x

Ellison SL, Mendel JT, Scudder JM et al (2013) Galaxy pairs in the Sloan Digital Sky Survey—VII. The merger-luminous infrared galaxy connection. Mon Not R Astron Soc 430:3128–3141. https://doi.org/10.1093/mnras/sts546

Ferrari F, De Carvalho RR, Trevisan M (2015) Morfometryka—a new way of establishing morphological classification of galaxies. Astrophys J 814:55. https://doi.org/10.1088/0004-637X/814/1/55

Flamary R (2016) Astronomical image reconstruction with convolutional neural networks

Freeman PE, Izbicki R, Lee AB et al (2013) New image statistics for detecting disturbed galaxy morphologies at high redshift. Mon Not R Astron Soc 434:282–295. https://doi.org/10.1093/mnras/stt1016

Gauci A, Adami KZ, Abela J (2010) Machine learning for galaxy morphology classification. arXiv preprint arXiv:1005.0390

Graff P, Feroz F, Hobson MP, Lasenby A (2014) SKYNET: an efficient and robust neural network training tool for machine learning in astronomy. Mon Not R Astron Soc 441:1741–1759. https://doi.org/10.1093/mnras/stu642

Habibzadeh M, Krzy A, Fevens T (2014) Comparative study of feature selection for white blood cell differential counts in low resolution images. Artif Neural Netw Pattern Recognit. https://doi.org/10.1007/978-3-319-11656-3_20

Hancer E, Xue B, Karaboga D, Zhang M (2015) A binary ABC algorithm based on advanced similarity scheme for feature selection. Appl Soft Comput J 36:334–348. https://doi.org/10.1016/j.asoc.2015.07.023

Hosny KM (2007) Exact and fast computation of geometric moments for gray level images. Appl Math Comput 189:1214–1222. https://doi.org/10.1016/j.amc.2006.12.025

Hosny KM (2011) Image representation using accurate orthogonal Gegenbauer moments. Pattern Recognit Lett 32:795–804. https://doi.org/10.1016/j.patrec.2011.01.006

Hosny KM (2014) New set of Gegenbauer moment invariants for pattern recognition applications. Arab J Sci Eng 39:7097–7107. https://doi.org/10.1007/s13369-014-1336-8

Hubble E (1936) The luminosity function of nebulae. II. The luminosity function as indicated by residuals in velocity-magnitude relations. Astrophys J 84:270

Huertas-Company M, Rouan D, Tasca L et al (2008) A robust morphological classification of high-redshift galaxies using support vector machines on seeing limited images I. Method description. Astron Astrophys 478:971–980. https://doi.org/10.1051/0004-6361:20078625

Huertas-Company M, Tasca L, Rouan D et al (2009) A robust morphological classification of high-redshift galaxies using support vector machines on seeing limited images II. Quantifying morphological k-correction in the COSMOS field at $1 < z < 2$: K s band vs I band. Astron Astrophys 497:743–753. https://doi.org/10.1051/0004-6361/200811255

Huertas-Company M, Aguerri JAL, Bernardi M et al (2010) Revisiting the Hubble sequence in the SDSS DR7 spectroscopic sample: a publicly available bayesian automated classification. Astron Astrophys 157:1–13. https://doi.org/10.1051/0004-6361/201015735

Ji J, Pang W, Zheng Y et al (2015) A novel artificial bee colony based clustering algorithm for categorical data. PLoS ONE 10:1–17. https://doi.org/10.1371/journal.pone.0127125

Karaboga D, Akay B (2009) A comparative study of artificial bee colony algorithm. Appl Math Comput 214:108–132. https://doi.org/10.1016/j.amc.2009.03.090

Karaboga D, Ozturk C (2011) A novel clustering approach: artificial bee colony (ABC) algorithm. Appl Soft Comput J 11:652–657. https://doi.org/10.1016/j.asoc.2009.12.025

Kheirdastan S, Bazarghan M (2016) SDSS-DR12 bulk stellar spectral classification: artificial neural networks approach. Astrophys Space Sci. https://doi.org/10.1007/s10509-016-2880-3

Liao S, Chen J (2013) Object recognition with lower order Gegenbauer moments. Lect Notes Softw Eng 1:387–391. https://doi.org/10.7763/LNSE.2013.V1.82

Ma M, Liang J, Guo M et al (2011) SAR image segmentation based on artificial bee colony algorithm. Appl Soft Comput J 11:5205–5214. https://doi.org/10.1016/j.asoc.2011.05.039

Peng CY, Ho LC, Impey CD, Rix H-W (2002) Detailed structural decomposition of galaxy images. Astron J. https://doi.org/10.1086/340952

Polsterer KL, Gieseke F, Igel C (2015) Automatic galaxy classification via machine learning techniques: Parallelized rotation/flipping INvariant Kohonen maps (PINK). In: Astronomical data analysis software an systems XXIV (ADASS XXIV), vol 495, p 81

Salouan R, Safi S, Bouikhalene B (2015) Handwritten Arabic characters recognition using methods based on Racah, Gegenbauer, Hahn, Tchebychev and Orthogonal Fourier-Mellin moments. Int J Adv Sci Technol 78:13–28

Schutter A, Shamir L (2015) Galaxy morphology—an unsupervised machine learning approach. Astron Comput 12:60–66. https://doi.org/10.1016/j.ascom.2015.05.002

Selim IM, Abd El Aziz M (2017) Automated morphological classification of galaxies based on projection gradient nonnegative matrix factorization algorithm. Exp Astron. https://doi.org/10.1007/s10686-017-9524-7

Selim IM, Hendy YHM, Bendary R (2014) Physical galaxy pairs and their effects on star formation. arXiv preprint arXiv:1410.7445

Shamir L (2011) Ganalyzer: a tool for automatic galaxy image analysis. Astrophys J. https://doi.org/10.1088/0004-637X/736/2/141

Sparke SL, Gallagher SJ (2007) Galaxies in the universe an introduction, 2nd edn. Cambridge University Press, Cambridge

Storrie-Lombardi MC, Lahav O, Sodre L, Storrie-Lombardi L Jr. (1992) Morphological classification of galaxies utilizing neural networks. Mon Not R Astron Soc 259:8–12

Von Hippel T (1994) Automated classification of stellar spectra-I. Initial results with artificial neural networks. Mon Not R Astron Soc 269:97–104

Wang W, Mottershead JE (2013) Adaptive moment descriptors for full-field strain and displacement measurements. J Strain Anal Eng Des 48:16–35. https://doi.org/10.1177/0309324712460485

Wang K, Ping Z, Sheng Y (2016) Development of image invariant moments—a short overview. Chin Opt Lett 14:1–11. https://doi.org/10.3788/col201614.091001.1

Willett KW, Galloway MA, Bamford SP et al (2017) Galaxy Zoo: morphological classifications for 120 000 galaxies in HST legacy imaging. Mon Not R Astron Soc 464:4176–4203. https://doi.org/10.1093/mnras/stw2568