Machine Learning for Galaxy Morphology Classification

Adam Gauci[1], Kristian Zarb Adami[2], John Abela[1]
[1]*Department of Intelligent Computer Systems, Faculty of ICT, University of Malta*
[2]*Department of Physics, Faculty of Science, University of Malta*

Released 2010 Xxxxx XX

ABSTRACT
In this work, decision tree learning algorithms and fuzzy inferencing systems are applied for galaxy morphology classification. In particular, the CART, the C4.5, the Random Forest and fuzzy logic algorithms are studied and reliable classifiers are developed to distinguish between spiral galaxies, elliptical galaxies or star/unknown galactic objects. Morphology information for the training and testing datasets is obtained from the Galaxy Zoo project while the corresponding photometric and spectra parameters are downloaded from the SDSS DR7 catalogue.

Key words: SDSS, Galaxy Zoo, galaxy morphology classification, decision trees, fuzzy logic, machine learning

1 INTRODUCTION

The exponential rise in the amount of available astronomical data has and will continue to create a digital world in which extracting new and useful information from archives is and will continue to be a major endeavour in itself. For instance, the seventh data release of the Sloan Digital Sky Survey (SDSS DR7), catalogues five-band photometry for 357 million distinct objects covering an area of 11,663 $\deg^2$, over 1.6 million spectra for 930,000 galaxies, 120,000 quasars, and 460,000 stars over 9380 $\deg^2$ (Abazajian 2009). Future wide-field imaging surveys will capture invaluable images for hundreds of millions of objects even those with very faint magnitudes.

Most of our current knowledge on galaxy classification is based on the pioneering work of several dedicated observers who visually catalogued thousands of galaxies. For instance, in Fukugita et al. (2007), 2253 objects from the SDSS DR3 were classified into a Hubble Type catalogue by three people independently. Then, a final classification was obtained from the mean. Classifying all objects captured in very large datasets produced by digital sky surveys is beyond the capacity of a small number of individuals. This therefore calls for new approaches. The challenge here is to design intelligent algorithms which will reproduce classification to the same degree as that made by human experts. Automated methods which make use of artificial neural networks have already been proposed by Ball et al. (2009) and more recently by Banerji et al. (2009). Andrae and Melchior also propose the use of shapelet decomposition to model the natural galaxy morphologies.

In this study, the advantages obtained by performing decision tree learning and fuzzy inferencing for galaxy mor-

phology classification are investigated. In particular, results from the CART, the C4.5 and the Random Forest tree building algorithms are compared. The outputs obtained after testing a generated fuzzy inference system through subtractive clustering, are also presented. Ball et al. (2009) adopted decision tree approaches for star/galaxy classification. Calleja and Fuentes (2004) also attempted to construct classifiers from parameters outputted from processed images of galaxies. However, such work was only constructed on a limited number of attributes and samples.

The aim here is to develop reliable models to distinguish between spiral galaxies, elliptical galaxies or star/unknown galactic objects. This through photometric as well as spectra data. Classified training and testing samples are obtained from the Galaxy Zoo project while the corresponding parameters are downloaded from the SDSS DR7 catalogue. As discussed by Banerji et al. (2009), this and similar studies present us the unique opportunity to compare human classifications to those obtained through automated machine learning algorithms. Should the automated techniques prove to be as successful as the human techniques in separating astronomical objects into different morphological classes, considerable time and effort will be saved in future surveys whilst also ensuring uniformity in the classifications.

In the following Section, the Galaxy Zoo Catalogue is described. Section 3 explores the various decision tree algorithms while Section 4 attempts to introduce fuzzy inference systems. In Section 5, the SDSS photometric and spectra parameters used for classification are discussed. The results obtained and the overall conclusion are then given in Section 6 and Section 7 respectively.
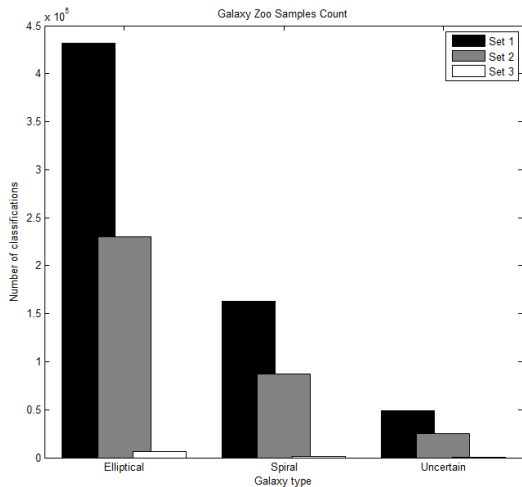
arXiv:1005.0390v2 [astro-ph.GA] 1 Jun 2010

Figure 1. Sample Counts

Figure 2. Decision tree for the concept of *ShuttleLaunch*

2   THE GALAXY ZOO DATASET

Galaxy classification is a task that humans excel in. Galaxy Zoo realises this and offers web users a valued service where volunteers can log in and classify galaxies according to their morphological class. The online portal (www.galaxyzoo.org) presents its users a sky image which centres on a galaxy randomly selected from a defined set. Such images are colour composites of the g, r and i bands available in the SDSS. Users are asked to determine whether the image shown consists of a spiral galaxy, an elliptical galaxy, a merger, a star or an unknown object. Users are also asked to distinguish between clockwise, anticlockwise or edge-on spirals. No distinction is made between barred or unbarred systems. This in turn creates a directory specifying the morphological class of each galaxy. Such a project has attracted a considerable number of users. An interesting weighting scheme which takes into account the similarity of the classifications of each user to the rest of the users, is adopted. Full details of this are available in Lintott et al. (2008).

In this work, use of the full sample set of Bamford et al. (2009), which is updated to include redshifts from SDSS DR7 was used. This is based on the magnitude-limited SDSS Main Galaxy Sample (Strauss et al. 2002). The dataset consists of 667,935 entries each of which corresponds to an object in the SDSS database (Set 1). For training and testing, debiased samples with a probability greater than 0.8 were considered (Set 2). In (Banerji et al. 2009), the same threshold was used on the raw morphological type likelihoods. Such discrimination is applied in order to identify the samples that were tagged with the same morphology by most users. After this filtering, the dataset reduced to 251,867 samples from which 75,000 were randomly selected to test the algorithms (Set 3). Using larger training and testing sets did not produce any gain in accuracy. Table 1 gives the number of galaxies in each morphological class for the defined sets whereas Figure 1 reproduces such data graphically. The trends in all sets are very similar. This indicates that each subset represents a true distribution of the samples in the entire catalogue.
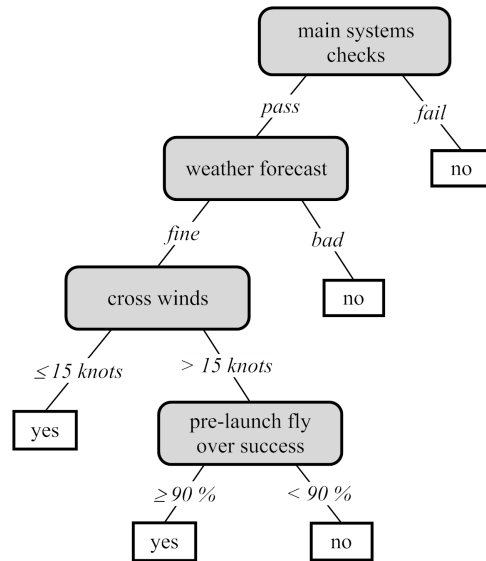
3   DECISION TREES

The task of classifying galaxies is a typical classification problem in which samples are to be categorised into groups from a discrete set of possible categories. The main objective is to create a model that is able to predict the class of a sample from several input parameters. Decision tree classification schemes sort samples by determining the corresponding leaf node after traversing down the tree from the root. A tree is learnt after splitting the examples into subsets based on an attribute value test. Branches of the tree are built by recursively repeating this process for each node and stops when all elements in the subset at a node have the same value of the target variable, or when splitting no longer adds value to the predictions. Some tree learning algorithms also allow for inequality tests and can work on noisy and incomplete datasets. Decision tree learning can be successfully applied when the samples can be described via a fixed number of parameters and when the output forms a discretised set. For instance, Figure 2 presents the learnt concept of *ShuttleLaunch* which suggests whether it is suitable to proceed with a scheduled launching after taking into account system checks and forecasted weather.

Like other supervised learning algorithms, a set of training examples with known classification is initially processed to infer a decision tree. Hopefully, this would be a good description of the classification procedure and will eventually be used to distinguish the class of other unseen samples. The learnt class description can be understood by humans and new knowledge about galaxy morphology classification in particular, can be obtained.

The approach adopted by most algorithms is very much the same. Moving in a top down manner, a greedy search that goes through the entire space of possible trees is initiated to try and find the minimum structure that correctly represents the examples. The constructed tree is then used as a rule set for predicting the class category of an unknown sample from the same set of attributes.

Table 1. Number of samples in each morphological class in each set

Set Name	Number of Objects	No. of Ellipticals	No. of Spirals	No. of Stars/Unknown objects
Set 1	667,935	431,436	230,207	6,292
Set 2	251,867	163,206	87,242	1,419
Set 3	75,000	49,193	25,273	534

3.1 C4.5

Following work done by Hunt in the late 1950s and early 1960s, Ross Quinlan continued to improve on the developed techniques and released the Iterative Dichottomizer 3 (ID3) and the improved C4.5 decision tree learners (Kohavi and Quinlan 1990). Even though the C5 algorithm is commercially available, the freely available C4.5 algorithm will be brought forward.

Here trees are built by recursively searching through and splitting the provided training set. If all samples in the set belong to the same class, the tree is taken to be made of just a leaf node. Otherwise, the values of the parameters are tested to determine a non trivial partition that separates the samples into the corresponding classes. In C4.5, the selected splitting criterion is the one that maximizes the information gain and the gain ratio.

Let $RF(C_j, S)$ be the relative frequency of samples in set S that belongs to class C_j. The information that identifies the class of a sample in set S is:

$$I(S) = -\sum_{j=1}^{x} RF(C_j, S) log(RF(C_j, S))$$

After applying a test T that separates set S in $S_1, S_2, ..., S_n$, the information gained is:

$$G(S, T) = I(S) - \sum_{i=1}^{t} \frac{|S_i|}{|S|} I(S_i)$$

The test that maximises $G(S, T)$ is selected at the respective node. The main problem with this approach is that it favours tests having a large number of outcomes such as those producing a lot of subsets each with few samples. Hence the gain ratio that also takes the potential information from the partition itself is introduced:

$$P(S, T) = -\sum_{i=1}^{t} \frac{|S_i|}{|S|} log\left(\frac{|S_i|}{|S|}\right)$$

If all samples are classified correctly, the tree may be overfitting the data and will fail when attempting to classify more general, unseen samples. Normally this is prevented by restricting some examples from being considered when building the tree or by pruning some of the branches after the tree is inferred. C4.5 adopts the latter strategy and remove some branches in a single bottom up pass.

One of the main advantages of the C4.5 algorithm is that it is capable of dealing with real, non-nominal attributes and so renders itself compatible with continuous parameters. It can also handle missing attribute data.

3.2 CART

The Classification and Regression Tree (CART) scheme was developed by Friedman and Breiman (Kohavi and Quinlan

1990). Although a divide and conquer search strategy similar to that of C4.5 is used, the resulting tree structure, the splitting criteria, the pruning method as well as the way missing values are handled, are redefined.

CART only allows for binary trees to be created. While this may simplify splitting and optimally partitions categorical attributes, there may be no good binary split for a parameter and inferior trees might be inferred. However, for multi-class problems, twoing may be used. This involves separating all samples in two mutually exclusive super-classes at each node and then apply the splitting criteria for a two class problem.

As a splitting criterion, CART uses the Gini diversity index. Let $RF(C_j, S)$ again be the relative frequency of samples in set S that belongs to class C_j, then the Gini index is defined as:

$$I_{gini}(S) = 1 - \sum_{j=1}^{x} RF(C_j, S)^2$$

and the information gain due to a particulat test T can be computed from:

$$G(S, T) = I_{gini}(S) - \sum_{i=1}^{t} \frac{|S_i|}{|S|} I_{gini}(S_i)$$

As with the C4.5 algorithm, the split that maximises $G(S, T)$, is selected. If all samples in a given node have the same parameter value, then the samples are perfectly homogenous and there is no impurity.

The CART algorithm also prunes the tree and use cross validation methods that may require more computation time. However, this will render shorter trees than those obtained from C4.5. Samples with missing data may also be processed.

3.3 Random Forests

Breiman and Cutler (2001), the pioneers of random forests, suggest using a classifier in which a number of decision trees are built. When processing a particular sample, the output by each of the individual trees is considered and the resulting mode is taken as the final classification. Each tree is grown from a different subset of examples allowing for an unseen (out of bag) set of samples to be used for evaluation. Attributes for each node are chosen randomly and the one which produces the highest level of learning is selected. It is shown that the overall accuracy is increased when the trees are less correlated. Having each of the individual trees with a low error rate, is also desirable.

Apart from producing a highly accurate classifier, such a scheme can also handle a very large amount of samples and input variables. A proximity matrix which shows how samples are related, is also generated. This is useful since such

relations may be very difficult to be detect just by inspection. With this strategy, good results may still be obtained even when a large portion of the data is missing or when the number of examples in each category is biased.

4 FUZZY INFERENCE SYSTEMS

Fuzzy logic accommodates soft computing by allowing for an imprecise representation of the real world. In crisp logic a clear boundary is considered to separate the various classes and each element is categorised into one group such that samples in sets A and $notA$ represents the entire dataset. Fuzzy logic extends on this by giving all sample a degree of membership in each set hence also caters for situations in which simple boolean logic is not enough. If classically set membership was denoted by 0 (false) or 1 (true), now we can also have 0.25 or 0.75. In fuzzy logic, the truth of any statement becomes a matter of degree.

The mathematical function that maps each input to the corresponding membership value between 0 and 1 is known as the membership function. Although this can be arbitrary, such function is normally chosen with computation efficiency and simplicity kept in mind. Various common membership functions include the triangular function, the trapezoidal function, the gaussian function and the bell function. The latter are the most popular and although they are smooth, concise and can attain non-zero values anywhere, they fail in specifying asymmetric membership functions. Such limitation is elevated through the use of sigmoid functions that can either open left or right.

For an inference system, $if-then$ rules that deal with fuzzy consequents and fuzzy antecedents are defined. An aggraded fuzzy set is then outputted after these conditional rules are compared and combined by standard logical operators equivalents. Since the degree of membership can now attain any value between 0 and 1, the AND and OR operators are replaced by the max and min functions respectively. The resulting output is then defuzzified to obtain one output value.

Fuzzy inference systems are easily understood and can even be applied when dealing with imprecise data. Like decision tree classifiers, they provide a penetrative model that experts can analyze and even add other information to it. Such inference approaches have already been successfully applied in a number of applications that range from integration in consumer produces to industrial process control, medical instrumentation and decision support systems.

5 INPUT PARAMETERS

In all machine learning algorithms, the set of input parameters strongly determine the overall accuracy of the classifier. Ideally, a minimum number of attributes that can differentiate between the three galaxy morphology classes are required. For this work, photometric and spectra values downloaded from the SDSS `PhotoObjAll` and `SpecLineAll` tables, were used. Data for which classification information is available in the Galaxy Zoo catalogue were downloaded and used to test the various machine learning algorithms used.

Table 2. Set of input parameters that are band independent, from the i band ($\approx 700nm - 1400nm$) and from the r band ($\approx 700nm$)

Name	Description
`dered_g - dered_r`	deredded (g - r) colour
`dered_r - dered_i`	deredded (r - i) colour
`deVAB_i`	DeVaucouleurs fit axis ratio
`expAB_i`	Exponential fit axis ratio
`lnLExp_i`	Exponential disk fit log likelihood
`lnLDeV_i`	DeVaucouleurs fit log likelihood
`lnLStar_i`	Star log likelihood
`petroR90_i / petroR50_i`	Concentration
`mRrCc_i`	Adaptive (+) shape measure
`texture_i`	Texture parameter
`mE1_i`	Adaptive E1 shape measure
`mE2_i`	Adaptive E2 shape measure
`mCr4_i`	Adaptive fourth moment
`deVAB_r`	DeVaucouleurs fit axis ratio
`expAB_r`	Exponential fit axis ratio
`lnLExp_r`	Exponential disk fit log likelihood
`lnLDeV_r`	DeVaucouleurs fit log likelihood
`lnLStar_r`	Star log likelihood
`petroR90_r / petroR50_r`	Concentration
`mRrCc_r`	Adaptive (+) shape measure
`texture_r`	Texture parameter
`mE1_r`	Adaptive E1 shape measure
`mE2_r`	Adaptive E2 shape measure
`mCr4_r`	Adaptive fourth moment

5.1 Photometric Attributes

In this study, the set of 13 parameters as taken by Banerji et al. (2009) which are based on colour, profile fitting and adaptive moments were used. However, we did not limit the evaluation to the i band but also aimed at testing whether the values derived from the r band give equal or better classification accuracies. The input parameters used are presented in Table 2.

The DeVaucouleurs law provides a measure of how the surface brightness of an elliptical galaxy varies with apparent distance from the centre. This should provide a good element of discrimination between spiral and elliptical profiles. The `lnLStar` parameter also helps to separate galaxy from star objects. The concentration parameter is given by the ratios of radii containing 90% and 50% of the Petrosian flux in a given band. The `texture` parameter compares the range of fluctuations in the surface brightness of the object to the full dynamic range of the surface brightness. It is expected that this is negligible for smooth profiles but becomes significant in high variance regions such as spiral arms.

The other parameters used are based on the object's shape. Particularly, the adaptive moments derived from the SDSS photometric pipeline are second moments of the object intensity, measured using a particular scheme designed to have an optimal signal to noise ratio. These moments are calculated by using a radial weight function that adopts to the shape and size of the object. Although theoretically there exists an optimal radial shape for the wight function related to the light profile of the object, a Gaussian with size matched to that of the object is used (Bernstein and Jarvis 2002).

The sum of the second moments in the CCD row and column direction (`mRrCc`) is calculated by:

$$mRrCc = <c^2> + <r^2>$$

where c and r correspond to the columns and rows of the sensor respectively and the second moments are defined as

$$<c^2> = \frac{\sum[I(r,c)w(r,c)c^2]}{\sum[I(r,c)w(r,c)]}$$

I is the intensity of the object and w is the weighting function. The ellipticity/polarisation components are defined by:

$$m_{e1} = <c^2> - \frac{<r^2>}{MRrCc}$$

$$m_{e2} = 2\frac{<(c)(r)>}{MRrCc}$$

A fourth order moment is also defined as:

$$m_{cr4} = \frac{<(c^2+r^2)^2>}{\sigma^4}$$

In this case, σ is the weight of the Gaussian function applied.

5.2 Spectra Attributes

To try and differentiate between elliptical, spiral and other morphologically shaped galaxies, this study also makes use of strong emission lines captured in galactic spectra. Significant lines of oxygen and hydrogen around the 5000Åand 7000Åmarks are expected for spiral galaxies. Since these often have star forming regions in the arms, the presence of sulphur, nitrogen and helium originating from ionized gas clouds, is also expected. Elliptical galaxies on the other hand are believed to have no star forming activity and can therefore be identified from continuous and average spectra.

Part of the SDSS pipeline is responsible to detect and store all strong emission lines present in the captured spectra. This is achieved through wavelet filters. An attempt to match all peaks with one of the candidate emission lines defined in a list, is made. Each line is then fitted with a single Gaussian by the SLATEC common mathematical library routine SNLS1E (SDSS). The height and the dispersion of the fitted Gaussian, the resulting Chi-squared error and other derived parameters are stored in the corresponding database. Table 3 presents the wavelengths of lines considered for this study. Lines storing missing dummy values for more than 5% of the dataset, were ignored. Although still randomly selected, samples for the training and testing sets were biased towards entries with a lower chi-squared error. This allowed for height values of better fitted Gaussians to be considered.

6 RESULTS

Initially, the 13 photometric parameters derived from the i band were standardised and independent component analysis was performed to determine the most significant components. As can be seen from the resulting eigenvalues shown in Figure 3, all of the independent components attain a non-zero value. This implies that all attributes are important for galaxy classification and dimension reduction is unnecessary. Figure 4 and Figure 5 show the eigenvalues obtained when

Table 3. Wavelengths of spectra lines

| Wave	Label	Wave	Label
3727.09	OII 3727	4960.30	OIII 4960
3729.88	OII 3730	5008.24	OIII 5008
3798.98	Hh 3799	5176.70	Mg 5177
3836.47	Oy 3836	5895.60	Na 5896
3889.00	HeI 3889	6302.05	OI 6302
3934.78	K 3935	6365.54	OI 6366
4072.30	SII 4072	6549.86	NII 6550
4102.89	Hd 4103	6564.61	Ha 6565
4305.61	G 4306	6585.27	NII 6585
4341.68	Hg 4342	6718.29	SII 6718
4364.44	OIII 4364	6732.67	SII 6733
4862.68	Hb 4863		
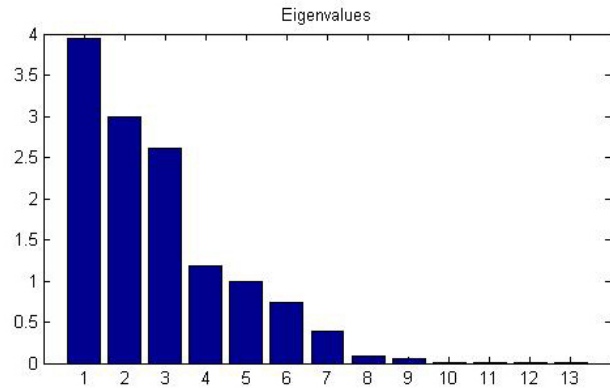
Figure 3. Eigenvalues from the 13 i band parameters

the same analysis was repeated on r band data as well as on the 24 attributes obtained when the i and r bands parameters were combined.

Once the significance of the selected parameters was confirmed, various machine learning algorithms were tested. In particular, the CART algorithm, the C4.5 algorithm with confidence values of 0.25 and 0.1 and the Random Forest algorithm with 10 and 50 trees were considered. For all test cases, a ten-fold cross validation strategy was used. The compiled 75,000 sample set (Set 3) was divided into 10 complimentary subsets and the learning algorithm was executed for 10 times. In each run, one of the ten subsets was used
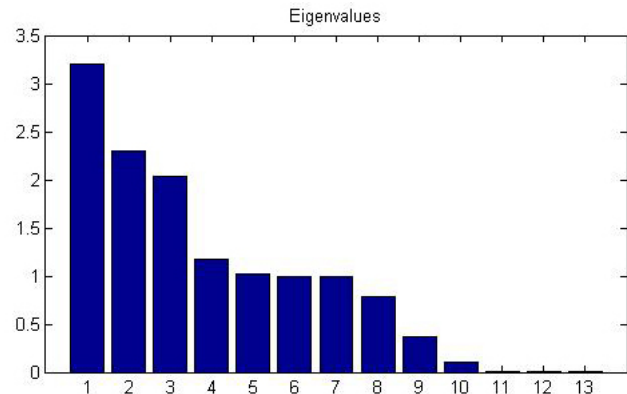

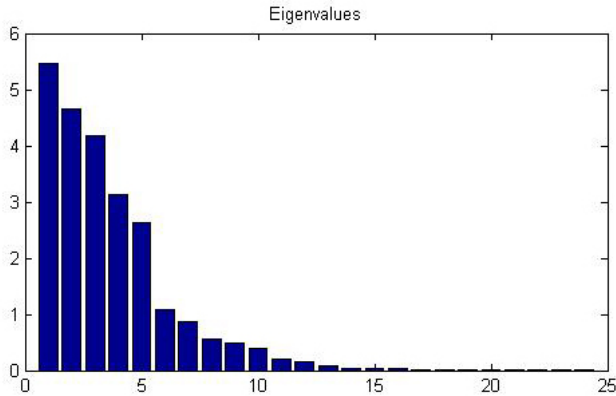**Figure 4.** Eigenvalues from the 13 r band parameters

Figure 5. Eigenvalues from the 24 i and r band parameters

Figure 6. Decision tree confusion matrices for the i band input parameters

for testing and the other nine subsets were put together to form the training set. The presented results are the computed averages across all ten trials. By this approach, every sample is part of the test set at least once.

6.1 i Band Photometric Parameters

The resulting confusion matrices when considering the 13 i band parameters are shown in Figure 6. These tables correlate the actual morphological classes with those outputted by the classifier. For instance, in the first row, the percentages of elliptical galaxies that were classified as elliptical (E), spiral (S) and unknown (U) are shown. Decision trees output a single morphology type for every input, therefore the percentages in each row add up to 100%. This corresponds to all of the input samples of a particular class. The global accuracy percentage was then calculated by comparing the total number of correctly classified samples with the total number of inputted tests.

In all decision tree algorithms tested, the global accuracy is always above 96.2% with the highest being 97.33% achieved by the 50 tree random forest technique. All confusion matrices result to have the highest values in the diagonal. This indicates that the majority of samples were classified correctly. The random forest algorithm with 50 trees did prove to be the most accurate and did manage to correctly classify 98.21% of all ellipticals, 96.10% of all spirals and 86.62% of all unknown objects. The slightly less than optimal classification percentages for unknown objects can be due to a number of factors. First of all, the number of training samples with unknown morphology might have not been enough for the algorithm to learn how to identify such samples and secondly, objects that mislead humans might actually have very similar properties to spiral or elliptical galaxies and are ultimately also classified correctly by the algorithm.

The membership functions derived by the fuzzy inference system for the DeVaucouleurs fit axis ratio, exponential fit axis ratio and concentration parameters in the i band, are shown in Figure 7. For such a model, subtracting clustering was used. The results obtained after testing are presented in Figure 8. Clearly, the developed model is capable of describing elliptical and spiral galaxies but suffers to accurately detect galaxies tagged to have an unknown type.
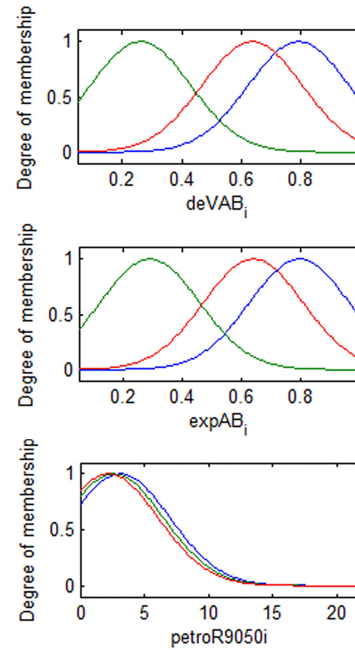

Figure 7. Fuzzy inference system membership functions

6.2 r Band Photometric Parameters

The CART, C4.5 and Random Forest decision tree algorithms were also applied to objects in Set 3 with the input parameters extracted from the r band. The resulting confusion matrices are presented in Figure 9. When compared to the results obtained from the i band data, a gain in the general accuracies of all algorithms can be noted. Using the

		FIS (Subtractive Clustering)		
		E	**S**	**U**
GZ	**E**	**96.68 %**	3.25 %	0.07 %
	S	7.85 %	**92.09 %**	0.06 %
	U	18.58 %	26.11 %	**55.31 %**

Classifier	**Accuracy**
FIS (Subtractive Clustering)	94.568 %

Figure 8. Fuzzy inference system confusion matrix for i band input parameters

CART

GZ	E	S	U
E	**97.53 %**	2.45 %	0.02 %
S	4.29 %	**95.51 %**	0.20 %
U	8.24 %	12.35 %	**79.41 %**

Random Forest (10 trees)

GZ	E	S	U
E	**98.44 %**	1.55 %	0.02 %
S	3.89 %	**96.02 %**	0.10 %
U	6.77 %	10.44 %	**82.79 %**

C4.5 (0.25 confidence)

GZ	E	S	U
E	**97.37 %**	2.57 %	0.05 %
S	4.25 %	**95.55 %**	0.20 %
U	6.62 %	11.03 %	**82.35 %**

Random Forest (50 trees)

GZ	E	S	U
E	**98.43 %**	1.55 %	0.02 %
S	3.23 %	**96.68 %**	0.10 %
U	6.32 %	9.12 %	**84.56 %**

C4.5 (0.1 confidence)

GZ	E	S	U
E	**97.45 %**	2.50 %	0.05 %
S	4.13 %	**95.66 %**	0.20 %
U	7.06 %	10.74 %	**82.21 %**

Classifier	Accuracy
CART	96.623 %
C4.5 (0.25 confidence)	96.567 %
C4.5 (0.1 confidence)	96.656 %
Random Forest (10 trees)	97.408 %
Random Forest (50 trees)	97.663 %

Figure 9. Decision tree confusion matrices for r band input parameters

CART

GZ	E	S	U
E	**97.50 %**	2.46 %	0.04 %
S	4.49 %	**95.31 %**	0.20 %
U	6.77 %	11.77 %	**81.47 %**

Random Forest (10 trees)

GZ	E	S	U
E	**98.40 %**	1.59 %	0.02 %
S	4.12 %	**95.78 %**	0.10 %
U	6.03 %	9.85 %	**84.12 %**

C4.5 (0.25 confidence)

GZ	E	S	U
E	**97.35 %**	2.62 %	0.03 %
S	4.28 %	**95.54 %**	0.18 %
U	7.50 %	10.74 %	**81.77 %**

Random Forest (50 trees)

GZ	E	S	U
E	**98.40 %**	1.59 %	0.02 %
S	3.21 %	**96.67 %**	0.11 %
U	4.71 %	7.79 %	**87.50 %**

C4.5 (0.1 confidence)

GZ	E	S	U
E	**97.47 %**	2.50 %	0.03 %
S	4.22 %	**95.60 %**	0.18 %
U	6.91 %	11.03 %	**82.06 %**

Classifier	Accuracy
CART	96.553 %
C4.5 (0.25 confidence)	96.540 %
C4.5 (0.1 confidence)	96.641 %
Random Forest (10 trees)	97.304 %
Random Forest (50 trees)	97.663 %

Figure 10. Decision tree confusion matrices for i and r bands input parameters

CART

GZ	E	S	U
E	**86.38 %**	13.52 %	0.10 %
S	8.78 %	**91.20 %**	0.02 %
U	16.87 %	29.72 %	**43.41 %**

Random Forest (10 trees)

GZ	E	S	U
E	**85.94 %**	13.82 %	0.24 %
S	8.16 %	**91.75 %**	0.10 %
U	12.06 %	21.91 %	**66.03 %**

C4.5 (0.25 confidence)

GZ	E	S	U
E	**85.14 %**	14.64 %	0.22 %
S	8.67 %	**91.29 %**	0.04 %
U	13.53 %	23.97 %	**92.50 %**

Random Forest (50 trees)

GZ	E	S	U
E	**86.38 %**	13.50 %	0.13 %
S	7.53 %	**92.42 %**	0.06 %
U	11.32 %	21.91 %	**66.77 %**

C4.5 (0.1 confidence)

GZ	E	S	U
E	**85.27 %**	14.51 %	0.21 %
S	8.40 %	**91.58 %**	0.03 %
U	13.38 %	24.27 %	**62.35 %**

Classifier	Accuracy
CART	89.082 %
C4.5 (0.25 confidence)	88.773 %
C4.5 (0.1 confidence)	89.000 %
Random Forest (10 trees)	89.385 %
Random Forest (50 trees)	89.968 %

Figure 11. Decision tree confusion matrices for spectra input parameters

	i band parameters	*r* band parameters	*i* and *r* bands parameters	Spectra parameters
CART	96.227 %	96.623 %	96.553 %	89.082 %
C4.5 (0.25 conf)	96.203 %	96.567 %	96.540 %	88.773 %
C4.5 (0.1 conf)	96.288 %	96.656 %	96.641 %	89.000 %
Random Forest (10 trees)	96.979 %	97.408 %	97.304 %	89.385 %
Random Forest (50 trees)	97.331 %	97.663 %	97.663 %	89.968 %
FIS	94.568 %			

Figure 12. Data, algorithms and results

r band information seems to help in distinguishing between ellipticals and spirals. However, the same cannot be said for the unclassified class since accuracies for this morphology class depreciated from an overall average of about 81% to that of about 75%.

6.3   *i* and *r* Bands Photometric Parameters

Following the tests described in Section 6.1 and Section 6.2 above, models built from the i and r bands parameters were tested. A 24 attribute dataset was constructed for objects defined in Set 3 and the same decision tree algorithms were applied to obtain corresponding classifiers. The results are presented in Figure 10. Although an improvement in accuracy is registered by the 50 tree Random Forest algorithm, this is only by a very small percentage.

6.4   Spectra Parameters

A similar methodology was adopted to test classification accuracies of models developed from spectra data. All wave line entries for objects in Set 3 were initially downloaded from the SDSS database. As described in Section 5.2, the wavelengths for which more than 95% of the data was available, were considered. This allowed for a 24 attribute feature space and the achieved results are presented in Figure 11.

Although still reasonably accurate, the general classification capability was found to be less than that obtained when photometric parameters were used. This could be due to the fact that peak spectral lines which are significant to detect spiral galaxies, are not always present.

7   CONCLUSION

In this study, accuracies for different galaxy morphology classification models developed through various machine learning techniques were obtained and analyzed. Results from the CART, the C4.5 and the Random Forest decision tree algorithms as well as the output from Fuzzy Inference Systems, were compared. The advantages gained by performing computations on different photometric parameters and on spectra attributes, were also investigated and put forward. Figure 12 serves as a good summary of which data and algorithms were used as well as the overall accuracies obtain. In all cases, the Random Forest gave the highest percentages especially when 50 trees were used.

All of the tested algorithms took only a few minutes to run on a normal personal computer. Although the presented results are for Set 3, experiments on Set 2 that stored more samples were also carried out. Accuracy percentages very close to the ones published, were obtained.
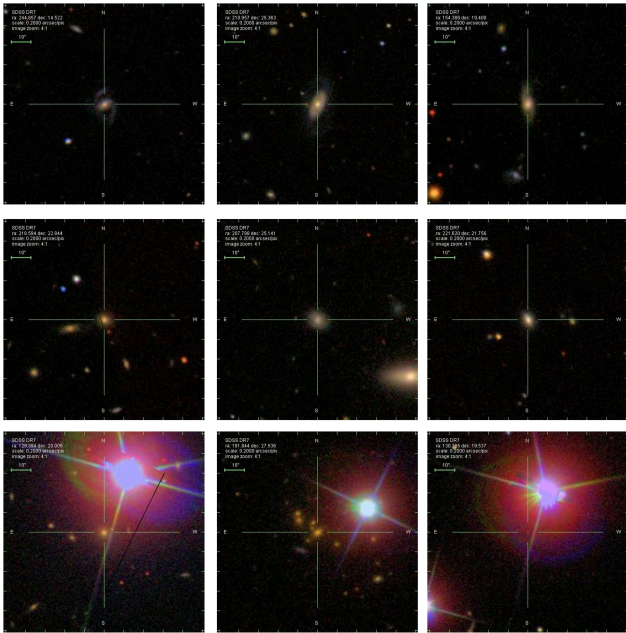
Figure 13. Samples of spiral (top), elliptical (middle) and unknown (bottom) galaxies that were incorrectly classified by the fuzzy inference system

In most cases, when processing photometric parameters, the adaptive shape measure (`mRrCc`) parameter was chosen as the root of the tree. First level nodes included the concentration (`petroR90`/`petroR50`) and the `dered_g-dered_r` parameters. For spectra data, the Ha wave line was determined to provide the highest information gain while the Hb and the K lines were chosen as first level nodes.

Figure 13 shows samples of incorrectly classified galaxies by the fuzzy inference system. Although this is not the most accurate technique described, the incorrectly classified spiral and elliptical samples are very faint in magnitude. Moreover, all incorrectly classified unknown objects have bright sources in the vicinity and this could have had an effect on the calculated parameters by the SDSS photometric pipeline.

8   ACKNOWLEDGEMENTS

REFERENCES

K. N. et al Abazajian. The seventh data release of the sloan sigital sky survey. *ApJS*, 182:543–558, 2009.

R. Andrae and P. Melchior. Morphological galaxy classification with shapelets.

N. M. Ball, J. Loveday, M. Fukugita, O. Nakamura, S. Okamura, J. Brinkmann, and R. J Brunner. Galaxy types in the sloan digital sky survey using supervised artificial neural networks. *MNRAS*, 348:1038–1046, 2009.

S. P. Bamford, R. C. Nichol, I. K. Baldry, K. Land, C. J. Lintott, K. Schawinski, A. Slosar, A. S. Szalay, D. Thomas, M. Torki, D. Andreescu, E. M. Edmondson, C. J. Miller, P. Murray, M. J. Raddick, and J. Vandenberg. Galaxy zoo: the dependence of morphology and colour on environment. *Monthly Notices of the Royal Astronomical Society*, 393:1324–1352, 2009.

M. Banerji, O. Lahav, C. J. Lintott, F. B. Abdalla, K. Schawinski, D. Andreescu, S. Bamford, P. Murray, M. J. Raddick, A. Slosar, A. Szalay, D. Thomas, and J. Vandenberg. Galaxy zoo: Reporducing galaxy morphologies via machine learning. 2009.

G. M. Bernstein and M. Jarvis. Shapes and shears, stars and smears: Optimal measurments for weak lensing. *The Astrophysical Journal*, 123:583–618, 2002.

L. Breiman and A. Cutler. Random forests, 2001.

J. Calleja and O. Fuentes. Automated classification of galaxy images. (3215):411–418, 2004.

M. Fukugita, O. Nakamura, S. Okamura, N. Yasuda, J. C. Barentine, J. Brinkmann, J. E. Gunn, M. Harvanek, T. Ichikawa, R. H. Lupton, D. P. Schneider, M. Strauss, and D. G. York. A catalog of morphologically classified galaxies from the sloan digital sky survey: North equatorial region. *The Astronomical Journal*, 134:579–593, 2007.

R. Kohavi and R. Quinlan. Decision tree discovery. 1990.

C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, P. Murray, and J. VanDenBerg. Galaxy goo: Morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *MNRAS*, 2008.

SDSS. Algorithms - emission and absorption line fitting.

M. A. Strauss, D. H. Weinberg, R. H. Lupton, V. K. Narayanan, J. Annis, M. Bernardi, M. Blanton, S. Burles, A. J. Connolly, J. Dalcanton, M. Doi, D. Eisenstein, J. A. Frieman, M. Fukugita, J. E. Gunn, Z. Ivezic, S. Kent, R. S. J. Kim, G. R. Knapp, R. G. Kron, J. A. Munn, H. J. Newberg, R. C. Nichol, S. Okamura, T. R. Quinn, D. J. Richmond, M. W.and Schlegel, K. Shimasaku, M. Sub-

baRao, A. S. Szalay, D. V. Berk, M. S. Vogeley, N. Yanny, B.and Yasuda, D. G. York, and I. Idit Zehavi. Spectroscopic target selection in the sloan digital sky survey: The main galaxy sample. *The Astronomical Journal*, 124: 1810–1824, 2002.