Identifying Uncertain Galaxy Morphologies using Unsupervised Learning

Kieran Jay Edwards and Mohamed Medhat Gaber

School of Computing, University of Portsmouth
Hampshire, England, PO1 3HE, UK
kieran.edwards@myport.ac.uk,mohamed.gaber@port.ac.uk

Abstract. With the onset of massive cosmological data collection through mediums such as the Sloan Digital Sky Survey (SDSS), galaxy classification has been accomplished for the most part with the help of citizen science communities like Galaxy Zoo. However, an analysis of one of the Galaxy Zoo morphological classification data sets has shown that a significant majority of all classified galaxies are, in fact, labelled as "Uncertain". This has driven us to conduct experiments with data obtained from the SDSS database using each galaxy's right ascension and declination values, together with the Galaxy Zoo morphology class label, and the k-means clustering algorithm. This paper identifies the best attributes for clustering using a heuristic approach and, accordingly, applies an unsupervised learning technique in order to improve the classification of galaxies labelled as "Uncertain" and increase the overall accuracies of such data clustering processes. Through this heuristic approach, it is observed that the accuracy of classes-to-clusters evaluation, by selecting the best combination of attributes via information gain, is further improved by approximately 10-15%. An accuracy of 82.627% was also achieved after conducting various experiments on the galaxies labelled as "Uncertain" and replacing them back into the original data set. It is concluded that a vast majority of these galaxies are, in fact, of spiral morphology with a small subset potentially consisting of stars, elliptical galaxies or galaxies of other morphological variants.

Keywords: Astronomical Data Mining, K-means, Cluster Identification, Classification Accuracy, Galaxy Morphology

1 Introduction

The fourth paradigm [1], to which it is now referred, describes the emergence of data mining within various scientific disciplines, including that of astronomy. The Sloan Digital Sky Survey [2] alone possesses, at present, over 1,000,000 galaxies, 30,000 stars and 100,000 quasars collated into several data sets. With such copious amounts of data being acquired from various astronomical surveys, it now becomes imperative that an automated model to processing this data be developed so as to be able to generate useful information. The goal of this approach is to then produce an outcome that will result in effective human

learning. It is the process of characterizing the known, assigning the new and discovering the unknown in such a data-intensive discipline that encompasses what astronomical data mining is all about [3].

Various classification techniques such as Nave Bayes [4, 5], C4.5 [6–8] and Artificial Neural Networks (ANN) [9] appear to be the more popular choices of methods when processing astronomical data. However, research carried out [10] involving calculating the Davies-Bouldin Validity Index (DBI) of the various attributes to determine the best combination for identifying correlations between morphological attributes and user-selected morphological classes motivated the direction of our research. A list of the top 10 attributes was presented and the best combinations of these, which produced the lowest DB Index values, were analyzed. It was ascertained that the larger the DBI value an attribute produced, the less useful it would be for clustering. Notably, these same attributes also proved less than useful in decision tree classification.

As an initial experiment, we obtained data for these 10 attributes from the Sloan Digital Sky Survey database for 2500 galaxies which were identified by their right ascension and declination through the Galaxy Zoo Classification data set. The k-means algorithm was then applied and evaluated using classes-to-clusters evaluation. However, despite using various subsets or combinations of these 10 attributes and re-clustering these sets reiteratively, the resulting accuracies never exceeded 55%. This encouraged us to then obtain 135 attributes from the SDSS database table from which the 10 originated, and apply a heuristic technique in order to find the best combination of those attributes with respect to their information gain levels.

In this paper, an investigation of how to select the best combination of attributes for clustering and determining the categories of these galaxies using an unsupervised approach is carried out. We also show that the heuristic technique applied to the attribute selection process, based on information gain levels, improves the classes-to-clusters accuracy by approximately 10-15%, thus further improving the classification of these galaxies.

The rest of the paper is organized as follows. Section 2 explores the various attempts at comparing algorithms and improving the classification process of astronomical data, while Section 3 details the clustering techniques, the focus mainly being on the k-means algorithm, which is used throughout this research. Section 4 provides a detailed discussion of the acquisition of the data sets; the pre-processing involved which includes the heuristic approach to attribute selection, and the unsupervised clustering experiments that were carried out. The results acquired from these experiments and the overall conclusion of this research together with a direction for future work is provided in Sections 5 and 6 respectively.

2 Related Work

A whole host of techniques and algorithms have already been applied to astronomical data sets with the goal of improving classification, including C4.5, Nave Bayes, Random Forest (RF) and Artificial Neural Networks (ANN).

ANNs are slowly being utilized more often as they have proven, with the correct training, to be an effective means of classification. One study [11], which took a supervised approach, trained an ANN to predict specific properties of galaxies, namely morphological classifications and redshifts, which proved reliable. Similarly, an ANN was trained [12] to identify broad absorption line quasars, a sub-class of active galactic nuclei. The results showed accuracies of approximately 92%. However, results of ANN implementations depend heavily on effective pre-processing. A comparison of different ANN algorithms [13] used to classify astronomical objects, also utilizing supervised learning, showed that the pre-processing methods used were insufficient and had an equally negative impact on both the algorithms that were being compared. Another issue with ANN algorithms is their inferiority when used with high dimensional data [15].

The Nave Bayes algorithm is probabilistic [14] in that it assumes that all attributes are statistically independent. As such, it works by assigning to each object it encounters the most probable target value. However, based on a classification comparison done [6] between the Nave Bayes algorithm, the RF algorithm and the C4.5 algorithm, it is observed that the Nave Bayes classification results, with its best accuracy of 43.62%, are nowhere near as robust as that of the RF algorithm's.

Classifying galaxies using the C4.5 algorithm to generate decision trees has also been the focus of much research. The main benefit of the C4.5 algorithm is the fact that it is efficient when dealing with numerical and nominal attributes alike. It has been applied [7] successfully in distinguishing between spiral and elliptical galaxies. It is noted that the global accuracy obtained from all C4.5 algorithm experiments consistently stayed above 96.2%. This involved using confidence levels of 0.1 as well as 0.25.

Random forests are also deemed suitable for dealing with astronomical data sets as they are designed for effective use with very large amounts of data. When used to classify stars, quasars and galaxies [15], the overall accuracy approximates between 91-95%, showing a significant improvement over the use of single tree classifiers such as the C4.5 algorithm. Classifying galactic images using RF [6] has also shown to outperform its C4.5 counterpart and the Nave Bayes algorithm.

3 Background

Clustering, in its unsupervised form, has always been one of the key areas in exploratory data analysis (e.g. astronomy). It is referred to [5], more commonly in astronomy circles, as "spatial clustering" or "angular clustering" on the sky. One of the main advantages of utilizing such a technique is the ability to discover

hidden clusters or structures in the presented data sets. Clustering, in a general sense, is the process of partitioning a data set into groups based on similarity between attributes of objects. A clustering algorithm is considered consistent [16] if it outputs an effectively-defined partition. The issues that plague the various clustering techniques include the determination of the quality of these partitions and the consistency of the overall results. The three subsections that follow will briefly describe a few of the various clustering techniques that have been developed, provide a more in-depth view of the k-means algorithm, and briefly describe cluster-to-class evaluation.

3.1 Clustering Techniques

EM (Expectation-Maximization) Clustering Algorithm - The EM algorithm is often used for clustering and is known especially for its ability to handle missing data. It is defined as an iterative means of calculating the maximum likelihood of parameters. With each cycle, there is an alternation between the expectation (E) and maximization (M) steps in which new parameter estimates, proven to not decrease the log-likelihood, are output. This process is reiterated until the best-fit maximum-likelihood solution over the initial model parameters is found [17].

Hierarchical Clustering Algorithm - The hierarchical clustering algorithm builds a hierarchy of clusters for analysis that can be achieved through either an agglomerative [18] or divisive [19] strategy (i.e. "bottom up" or "top down" approach respectively). Some of the choices of metrics, depending on the nature of the objective of clustering, can be the Euclidean distance, Manhattan distance, maximum distance or even cosine similarity. The advantage that this algorithm has lies in its ability to use any measure of distance so long as it is valid.

Spectral Clustering Algorithm - Spectral clustering techniques are used to solve problems in graph partitions where different measures require optimization. This involves a two-step process [20]: Taking the various data points from more "obvious" clusters and embedding them in a space, followed by the application of a classical clustering algorithm such as the k-means algorithm.

3.2 K-means Algorithm

The k-means algorithm is one of the most popular clustering techniques available, used extensively in both industrial and scientific applications for cluster analysis. It is known [21] as a partitional or nonhierarchical clustering technique in which the aim is to partition n objects into k clusters where each object belongs to the cluster with the closest mean. This is an iterative, heuristic approach which starts with the assignment for each object as described in equation (1) given an initial set of k means $m_1^{(1)}, \ldots, m_k^{(1)}$, where each x_j is allocated to one $S_i^{(t)}$.

$$S_i^{(t)} = \{x_j : ||x_j - m_i^{(t)}|| \leq ||x_j - m_{i^*}^{(t)}|| \forall i^* = 1, \ldots, k\} \tag{1}$$

Table 1. Galaxy Zoo Table 2 Data Set: Final Morphological Classifications

Category	No. of Galaxies
Uncertain	41556
Spiral	17747
Elliptical	6232

This is followed by the calculation of the new means which is to become the newly appointed centroid of the cluster as shown in equation (2).

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \tag{2}$$

The iteration of these two steps will continue until convergence is achieved. When this occurs, the assignments of the centroids no longer change. The number of iterations required to achieve convergence can vary greatly which makes this algorithm potentially computationally intensive particularly with very large data sets. However, there are a number of variants of the k-means algorithm which address this problem [22, 23], improving its efficiency.

3.3 Classes-to-Clusters Evaluation

In the various experiments carried out, we take an unsupervised approach by using the k-means algorithm together with classes-to-clusters evaluation in order to evaluate the resulting clustering of the data and determine its accuracy. In classes-to-clusters evaluation, the class label (i.e. Spiral, Elliptical and Uncertain) is first ignored and the clusters generated using only numerical data. The clusters are then assigned classes based on the largest number of objects in each cluster that fall into a certain class. A classification error and confusion matrix are then computed which shows the resulting accuracy of the process.

4 Methodology

We started by obtaining the galaxy morphological classification voting data from the Galaxy Zoo Table 2 [24] data set and observed that a significant majority of all galaxies, approximately 63%, have been classified as "Uncertain". Table 1 shows the final classification result.

This led us to obtain data for the 10 attributes that were calculated to have the lowest DBI index values [10] from the SDSS database [25]. The three flag attributes used in the Galaxy Zoo Table 2 data set to indicate the morphology of the galaxies were combined, in order to decrease sparseness, into one attribute labeled "CLASS" and then included together with the 10 attributes. Table 2 shows the list of the 10 attributes used for this initial experiment.

The relational algebraic query used to retrieve data from the SDSS database to obtain the data for the 10 attributes can be expressed as follows.

Table 2. The 10 Attributes with the Lowest DBI Index Values

Attribute	Description
*isoAGrad u*z*	Gradient of the isophotal major axis
*petroRad u*z*	Petrosian radius
texture u	Measurement of surface texture
*isoA z*z*	Isophotal major axis
lnLExp u	Log-likelihood of exponential profile fit (typical for a spiral galaxy)
lnLExp g	Log-likelihood of exponential profile fit (typical for a spiral galaxy)
*isoA u*z*	Isophotal major axis
*isoB z*z*	Isophotal minor axis
*isoBGrad u*z*	Gradient of the isophotal minor axis
*isoAGrad z*z*	Gradient of the isophotal major axis

Table 3. The Best Resulting Subset of the Original 10 Attributes

Attribute
*isoA z*z*
lnLExp g
*isoAGrad u*z*
*isoB z*z*

result = $\pi\ \sigma\ isoAGrad_u{*}z/a.isoAGrad_u{*}z, \sigma\ petroRad_u{*}z/a.petroRad_u{*}z, \sigma$a.texture_u, $\sigma\ isoA\_z{*}z/a.isoA\_z{*}z, \sigma\ lnLExp_u/a.lnLExp\_u, \sigma\ lnLExp\_g/a.lnLExp\_g, \sigma\ isoA\_u{*}z/a.isoA\_u{*}z, \sigma\ isoB\_z{*}z/a.isoB\_z{*}z, \sigma\ isoBGrad\_u{*}z/a.isoBGrad\_u{*}z, \sigma\ isoAGrad\_z * z/a.isoAGrad\_z * z\ (u.up\_id = x.up\_id \wedge$ x.objID=p.objID $\wedge$ p.objID=a.objID ($x(\#x)$ |×| $u(\#upload)$ |×| p(PhotoTag) |×| a(PhotoObjAll)))

Accurate application of the morphological class labels to each of the galaxies in the data set before clustering was achieved by reference to each galaxy's right ascension and declination values. These produced, in the SDSS database query, the object ID for each galaxy which was then matched up to the object ID in the Galaxy Zoo Table 2 data set to obtain the correct label (i.e. Spiral, Elliptical or Uncertain). The k-means clustering algorithm was applied to the full data set of 3000 galaxies using classes-to-clusters evaluation with the value of k set to 3. This process was then repeated reiteratively using various subsets of the 10 attributes. The best resulting subset is shown in table 3.

The following R code used to apply the k-means algorithm to the data set and compare the clusters to classes.

```
Library(RWeka)
sdssTable <- read.csv(file=sdss.csv)
sdssTable2 <- sdssTable
results <- SimpleKMeans(sdssTable2[,-5], Weka_control(N=3))
results
table(predict(results), sdssTable\$CLASS)
```

Table 4. The Best Combination of Attributes with Respective Information Gain Levels

Attribute	Information Gain	Attribute	Information Gain
expRad_g	0.2207	*isoAGrad_r*	0.0775
expRad_r	0.1965	*lnLDeV_z*	0.0716
expRad_i	0.1831	*texture_g*	0.0706
lnLDeV_g	0.1367	*isoPhiGrad_g*	0.0639
lnLDeV_r	0.1275	*texture_r*	0.0522
isoB_i	0.1206	*lnLDeV_u*	0.0428
isoB_r	0.1154	*texture_i*	0.0367
lnLExp_r	0.1002	*isoPhiGrad_i*	0.03
lnLExp_i	0.0986	*texture_u*	0.0153
isoBGrad_g	0.092	*isoColcGrad_r*	0.0115
petroRad_u	0.0834		
lnLExp_z	0.0822		

It was originally thought that the reason for the low accuracies was due to the majority of the galaxies having been labeled as "Uncertain". An alternative clustering attempt, where 1000 of the 1763 galaxies labelled as "Uncertain" were removed, was carried out but proved ineffective as it showed no improvement whatsoever. In fact, the accuracy level dropped even further.

The objective of this paper is to be able to provide astronomers with a tool to effectively assign each galaxy to the right category as accurately as possible. With that in mind, we decided to re-query the SDSS database, this time obtaining 135 attributes, and apply a heuristic technique in order to obtain the best combination. This was achieved through the use of each of the attribute's information gain levels. Refer to the appendix for the full relational algebraic query that was used for data retrieval from the SDSS database.

4.1 Best Attribute Combination through Information Gain

After acquiring the 135 attributes for 5000 galaxies and pre-processing the data set, which involved removing selected attributes and objects that contained significant numbers of entries with the value -9999, the information gain level for all attributes was calculated and then listed in descending order. The heuristic technique was then employed. This involved clustering the data with the single attribute that possessed the highest information gain level together with the class label, using clusters-to-classes evaluation with the value of k set to 3. Once this 1st iteration completed, the attribute with the 2nd highest information gain level was then added in and the data re-clustered. If the accuracy level decreased, that 2nd attribute would then be removed and then the 3rd added. If the accuracy level remained the same or increased, that attribute would remain and the next attribute added on.

The final data set contained 4979 galaxies and 23 attributes. Table 4 lists the attributes and their information gain levels.

Table 5. The Iterative Clustering Results of the 10 attributes

No. of Attributes	Accuracy (%)	Within Cluster Sum of Squared Errors
1	50.8	0.31097208092561296
2	54.2	73.22356287236981
3	54.2	3.213059611539209
4	54.2	56.57163126388849
4	49.5333	3.213059611539209
5	49.4	5.1038948660063035
10	45.8	186.893316665896

4.2 Hidden Cluster Discovery and Labeling by Unsupervised Clustering

After acquiring the best combination of attributes, various clustering experiments were carried out in an attempt to accurately classify the galaxies labelled as "Uncertain". In order to further analyze these galaxies, we split them into two clusters: "cluster0" and "cluster1". These were saved and placed back into the original data set. The "cluster0" and "cluster1" clusters were then re-labeled; starting with "cluster0" being re-labeled as "spiral" and "cluster1" as "elliptical", and then the data set was clustered with the value of k set to 2. The labels of "cluster0" and "cluster1" were then reversed and the process repeated.

5 Experimental Results

The results of the initial experiments done on the 10 attributes with the lowest *DBI Index* Values [10] with 3000 galaxies and the value of k set to 3 are shown in Table 5.

After the 4th attempt, the accuracy lowered whenever additional attributes were added so the best subset contained only 4 attributes in the end.

After utilizing our heuristic technique to obtain the best selection from the 135 attributes obtained from the SDSS database, we conducted several clustering experiments on our final data set consisting of 4979 galaxies and 23 attributes which involved separating clusters, re-clustering, re-labeling and re-combining them together. Table 6 shows the various k-means clustering results accordingly.

It is notable that the highest clustering accuracy of 82.627% was obtained when galaxies from both "cluster0" and "cluster1" were re-labelled as "Spiral" galaxies. Out of the 4979 galaxies in the complete data set, only 865 were incorrectly classified. Motivated by this boost in accuracy, we conducted another set of experiments using state-of-the-art classification techniques, namely Random Forest (RF) [26] and Support Vector Machines (SVM) [27]. Table 7 lists the results of these additional experiments.

The accuracies for all three algorithms, when all the galaxies from "cluster0" and "cluster1" are re-labelled as "Spiral", consistently outperform the rest of the experiments. With the number of trees set to 100, Random Forest provided an

Table 6. The Results of the various k-means Clustering Experiments

Data Set Type	Number of Galaxies Per Cluster			Accuracy (%)
	Spiral	Elliptical	Uncertain	
Full Data Set	1476	520	2983	65.6156
Spiral/Elliptical Only	1476	520	-	72.495
Uncertain Only	-	-	2983	78.9474
Cluster0 - Spiral / Cluster1 - Elliptical	2104	2875	-	63.0649
Cluster0 - Elliptical / Cluster1 - Spiral	3831	1148	-	77.2444
Cluster0 - Spiral / Cluster1 - Spiral	4459	520	-	82.627
Cluster0 - Elliptical / Cluster1 - Elliptical	1476	3503	-	68.4475

Table 7. The Additional Experiments Involving RF and SVM Classification Techniques

Data Set Type	Algorithm Accuracy (%)		
	k-means	RF	SVM
Cluster0 - Spiral / Cluster1 - Elliptical	63.0649	90.6005	86.9452
Cluster0 - Elliptical / Cluster1 - Spiral	77.2444	83.6513	77.9675
Cluster0 - Spiral / Cluster1 - Spiral	82.627	91.3838	89.6566
Cluster0 - Elliptical / Cluster1 - Elliptical	68.4475	83.089	78.3892

exceptional accuracy of 91.3838% which indicates two concluding remarks that we can state with certainty:

- A significant majority of the galaxies labelled as "Uncertain" are indisputably of spiral morphology.
- There is another small subset of galaxies amongst those that are "Uncertain" that are either of elliptical morphology, are stars or possess an entirely different morphology type.

6 Conclusion & Future Work

Motivated by the fact that 60% of all galaxies in the Galaxy Zoo Table 2 data set are classified as "Uncertain", we attempted to introduce a means for astronomers to more efficiently and accurately classify these galaxies. We introduced a novel approach to accomplishing such a task by first utilizing a heuristic technique in order to obtain the best combination of attributes through their calculated information gain which serves to increase the clustering accuracy. We then conducted

a series of experiments involving the clustering of the galaxies labelled as "Uncertain", saving their cluster assignments and then re-introducing them back into the original data set. We have shown that the highest accuracy (82.627%) was obtained when all the galaxies from "cluster0" and "cluster1" were re-labelled as "Spiral" galaxies. Applying the Random Forest and SVM classification algorithms over all the original experiments further reinforced this finding. There is no doubt that a majority of the galaxies labelled as "Uncertain" in our data set are, in fact, of spiral morphology.

Avenues for future work include a large-scale processing of the right ascension and declination values for all 65535 galaxies and re-running the experiments that were performed in this paper.

References

1. Ball, N. M., and Brunner, R. J.: Data Mining and Machine Learning in Astronomy. In: International Journal of Modern Physics D (2009), pp. 61.
2. Stoughton, C., Lupton, R. H., Bernardi, M., Blanton, M. R., Burles, S., Castander, F. J., et al: Sloan Digital Sky Survey: Early Data Release. In: The Astronomical Journal 123.1 (2007), pp. 485.
3. Borne, K.: Scientific Data Mining in Astronomy. In: Next Generation of Data Mining (2009), pp. 91-114.
4. Henrion, M., Mortlock, D. J., Hand, D. J. and Gandy, A.: A Bayesian Approach to Star-Galaxy Classification. In: Monthly Notices of the Royal Astronomical Society (2011), pp. 2286-2302.
5. Kamar, E., Hacker, S. and Horvitz, E.: Combining Human and Machine Intelligence in Large-Scale Crowdsourcing. In: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (2012), pp. 467-474.
6. De La Calleja, J. and Fuentes, O.: Automated Classification of Galaxy Images. In: Knowledge-Based Intelligent Information and Engineering Systems (2004), pp. 411-418.
7. Gauci, A., Adami, K. Z. and Abela, J.: Machine Learning for Galaxy Morphology Classification. [arXiv:1005.0390] (2010), pp. 1-9.
8. Vasconcellos, E. C., de Carvalho, R. R., Gal, R. R., LaBarbera, F. L., Capelato, H. V., Velho, H. F. C., and Ruiz, R. S. R.: Decision Tree Classifiers for Star/Galaxy Separation. In: The Astronomical Journal 141 (2011), pp. 189.
9. Banerji, M., Lahav, O., Lintott, C. J., Abdalla, F. B., Schawinski, K., Bamford, S. P., Andreescu, D., Murray, P., Raddick, M. J., Slosar, A., Szalay, A, Thomas, D. and Vandenberg, J.: Galaxy Zoo: Reproducing Galaxy Morphologies Via Machine Learning. In: Monthly Notices of the Royal Astronomical Society (2010), pp. 342-353.
10. Baehr, S, Vedachalam, A., Borne, K. D. and Sponseller, D.: Data Mining the Galaxy Zoo Mergers. In: 2010 Conference on Intelligent Data Understanding (2010)
11. Ball, N. M., Loveday, J., Fukugita, M., Nakamura, O., Okamura, S., Brinkmann, J. and Brunner, R. J.: Galaxy Types in the Sloan Digital Sky Survey Using Supervised Artificial Neural Networks. In: Monthly Notices of the Royal Astronomical Society (2004), pp. 1038-1046.
12. Scaringi, S., Cottis, C. E., Knigge, C. and Goad, M. R.: Broad Absorption Line Quasar Catalogues with Supervised Neural Networks. [arXiv:0810.4396]. (2008)

13. Bazell, D. and Peng, Y.: A Comparison of Neural Network Algorithms and Preprocessing Methods for Star-Galaxy Discrimination. In: The Astrophysical Journal Supplement Series, vol. 116, no. 1, pp. 47 (2009)
14. Frank, E., Hall, M. and Pfahringer, B.: Locally Weighted Nave Bayes. In: Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence (2002), pp. 249-256.
15. Gao, D., zhang, y. X. and Zhao, Y. H.: Random Forest Algorithm for Classification of Multi-wavelength Data. In: Research in Astronomy and Astrophysics 9.2 (2009), pp. 220.
16. Von Luxburg, U., Bousquet, O. and Belkin, M.: Limits of Spectral Clustering. In: Advances in Neural Information Processing Systems (NIPS) (2005), pp. 857-864.
17. Bradley, P. S., Fayyad, U. and Reina, C.: Scaling EM (Expectation-Maximization) Clustering to Large Databases. In: Microsoft Research (1998)
18. Karypis, G., Han, E. H. and Kumar, V.: Chameleon: Hierarchical Clustering Using Dynamic Modeling. In: Computer, vol. 32, no. 8, pp. 68-75 (1999)
19. Ding, C. and He, X.: Cluster Merging and Splitting in Hierarchical Clustering Algorithms. In: Proceedings of the 2002 IEEE International Conference on Data Mining (2002), pp. 139-146.
20. Bengio, Y., Paiement, J. F., Vincent, P., Delalleau, O., Le Roux, N. and Ouiment, M.: Out-of-Sample Extensions for Lle, Isomap, Mds, Eigenmaps and Spectral Clustering. In: Advances in Neural Information Processing Systems 16 (2004), pp. 177-184.
21. Huang, Z.: Extensions to the K-means Algorithm for Clustering Large Data Sets with Categorical Values. In: Data Mining and Knowledge Discovery, vol. 2, no. 3, pp. 283-304 (1998)
22. Alsabti, K., Ranka, S. and Singh, V.: An Efficient K-Means Clustering Algorithm. In: Electrical Engineering and Computer Science, no. 43 (1997)
23. Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R. and Wu, A. Y.: An Efficient K-Means Clustering Algorithm: Analysis and Implementation. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 881-892 (2002)
24. Lintott C., Schawinski, K., Bamford, S., Slosar, A., Land, K., Thomas, D., et al: Galaxy Zoo 1: Data Release of Morphological Classifications for Nearly 900.000 Galaxies. In: Monthly Notices of the Royal Astronomical Society, vol. 410, no. 1, pp. 166-178 (2011)
25. Abazajian, K. N., Adelman-McCarthy, J. K., Agueros, M. A., Allam, S. S., Prieto, C. A., An, D., et al: The Seventh Data Release of the Sloan Digital Sky Survey. In: The Astrophysical Journal Supplement Series (2009), pp. 543.
26. Breiman, L.: Random Forests. In: Machine Learning, vol. 45, no. 1, pp. 5-32 (2001)
27. Cortes, C. and Vapnik, V.: Support-Vector Networks. In: Machine Learning, vol. 20, no. 3, pp. 273-297 (1995)

Appendix: SDSS Database Query (135 Attributes)

The relational algebraic query for the data of the 135 attributes that was used for data retrieval from the SDSS database.

$result = \pi$ a.petroMag_u, a.petroMag_g, a.petroMag_r, a.petroMag_i, a.petroMag_z, a.petroRad_u, a.petroRad_g, a.petroRad_r, a.petroRad_i, a.petroRad_z, a.petroR90_u, a.petroR90_g, a.petroR90_r, a.petroR90_i, a.petroR90_z, a.isoRowc_u, a.isoRowc_g,

a.isoRowc_r, a.isoRowc_i, a.isoRowc_z, a.isoRowcGrad_u, a.isoRowcGrad_g, a.isoRowcGrad_r, a.isoRowcGrad_i, a.isoRowcGrad_z, a.isoColc_u, a.isoColc_g, a.isoColc_r, a.isoColc_i, a.isoColc_z, a.isoColcGrad_u, a.isoColcGrad_g, a.isoColcGrad_r, a.isoColcGrad_i, a.isoColcGrad_z, a.isoA_u, a.isoA_g, a.isoA_r, a.isoA_i, a.isoA_z, a.isoB_u, a.isoB_g, a.isoB_r, a.isoB_i, a.isoB_z, a.isoAGrad_u, a.isoAGrad_g, a.isoAGrad_r, a.isoAGrad_i, a.isoAGrad_z, a.isoBGrad_u, a.isoBGrad_g, a.isoBGrad_r, a.isoBGrad_i, a.isoBGrad_z, a.isoPhi_u, a.isoPhi_g, a.isoPhi_r, a.isoPhi_i, a.isoPhi_z, a.isoPhiGrad_u, a.isoPhiGrad_g, a.isoPhiGrad_r, a.isoPhiGrad_i, a.isoPhiGrad_z, a.deVRad_u, a.deVRad_g, a.deVRad_r, a.deVRad_i, a.deVRad_z, a.deVAB_u, a.deVAB_g, a.deVAB_r, a.deVAB_i, a.deVAB_z, a.deVPhi_u, a.deVPhi_g, a.deVPhi_r, a.deVPhi_i, a.deVPhi_z, a.deVMag_u, a.deVMag_g, a.deVMag_r, a.deVMag_i, a.deVMag_z, a.expRad_u, a.expRad_g, a.expRad_r, a.expRad_i, a.expRad_z, a.expAB_u, a.expAB_g, a.expAB_r, a.expAB_i, a.expAB_z, a.expPhi_u, a.expPhi_g, a.expPhi_r, a.expPhi_i, a.expPhi_z, a.expMag_u, a.expMag_g, a.expMag_r, a.expMag_i, a.expMag_z, a.modelMag_u, a.modelMag_g, a.modelMag_r, a.modelMag_i, a.modelMag_z, a.texture_u, a.texture_g, a.texture_r, a.texture_i, a.texture_z, a.lnLExp_u, a.lnLExp_g, a.lnLExp_r, a.lnLExp_i, a.lnLExp_z, a.lnLDeV_u, a.lnLDeV_g, a.lnLDeV_r, a.lnLDeV_i, a.lnLDeV_z, a.fracDeV_u, a.fracDeV_g, a.fracDeV_r, a.fracDeV_i, a.fracDeV_z, a.dered_u, a.dered_g, a.dered_r, a.dered_i, a.dered_z(u.up_id=x.up_id ∧ x.objID=p.objID ∧ p.objID=a.objID (x(#x) | × | u(#upload) | × | p(PhotoTag) | × | a(PhotoObjAll)))