DATA MINING THE GALAXY ZOO MERGERS

STEVEN BAEHR*, ARUN VEDACHALAM*, KIRK BORNE*, AND DANIEL SPONSELLER*

Abstract. Collisions between pairs of galaxies usually end in the coalescence (merger) of the two galaxies. Collisions and mergers are rare phenomena, yet they may signal the ultimate fate of most galaxies, including our own Milky Way. With the onset of massive collection of astronomical data, a computerized and automated method will be necessary for identifying those colliding galaxies worthy of more detailed study. This project researches methods to accomplish that goal. Astronomical data from the Sloan Digital Sky Survey (SDSS) and human-provided classifications on merger status from the Galaxy Zoo project are combined and processed with machine learning algorithms. The goal is to determine indicators of merger status based solely on discovering those automated pipeline-generated attributes in the astronomical database that correlate most strongly with the patterns identified through visual inspection by the Galaxy Zoo volunteers. In the end, we aim to provide a new and improved automated procedure for classification of collisions and mergers in future petascale astronomical sky surveys. Both information gain analysis (via the C4.5 decision tree algorithm) and cluster analysis (via the Davies-Bouldin Index) are explored as techniques for finding the strongest correlations between human-identified patterns and existing database attributes. Galaxy attributes measured in the SDSS green waveband images are found to represent the most influential of the attributes for correct classification of collisions and mergers. Only a nominal information gain is noted in this research, however, there is a clear indication of which attributes contribute so that a direction for further study is apparent.

1. Introduction

1.1. **Scientific Rationale.** Current computational detection of a galaxy merger in astronomical data is less than ideal. However, human pattern recognition easily identifies mergers with varied, but strong, levels of accuracy. If this superior human input can be incorporated into the automated data pipeline detection scheme, informed by machine learning models, then a more accurate assessment of merger presence can be gained automatically in future large sky surveys. These improvements could potentially lead to more powerful detection of various astronomical objects and interactions.

Our goal was to generate merger classification models using two prominent machine learning approaches, as a preliminary exercise toward the incorporation of human input into future automated pipeline classification models.

1.2. **Citizen Science.** Citizen Science refers to the involvement of layperson volunteers in the science process, with the volunteers specifically asked to perform routine but authentic science research tasks that are beyond the capability of machines. Complex pattern recognition (and classification) and anomaly detection in complex data are among the types of tasks that would qualify as Citizen Science activities. The Galaxy Zoo project (galaxyzoo.org) presents imagery from the Sloan Digital Sky Survey (SDSS) to laypersons for classification (e.g., whether a galaxy is of the elliptical or spiral type) via a web interface. The project went live in 2007, and already over 200 million classifications have been provided by more than 260,000 individuals. During the classification process, volunteers can flag a particular image as depicting a merger of two or more galaxies. Approximately 3000 prominent mergers in the SDSS (Sloan Digital Sky Survey) have been identified[3].

*George Mason University, Fairfax, VA.

1.3. **Related Work.** Image recognition has long been a major deficiency in computation. Classification tasks such as facial recognition, trivially exercised with great accuracy and precision by living organisms, have been predominantly inaccurate and slow when attempted using computers. While current algorithms are fairly capable of recognizing substructures and details in imaging data, recognition of gestalt in the data has proved more elusive. This shortcoming, combined with the contemporary unyielding influx of data in the natural sciences and the vastness of a data domain such as astronomy, has led to the necessity of attempting to tap into the effortless capability of human cognition.

The Galaxy Zoo web application has as its goal the collection and application of human classifications applied to images of galaxies from the SDSS. Efforts have been made to use human input to reinforce existing machine learning models such as artificial neural networks and genetic algorithms[2]. Additionally, work has been done using supervised learning algorithms to classify galaxy type (non-merging), with considerable success using spectroscopic data for training[1] and data derived from human cognition[6]. It has been found that the introduction of parameters chosen using human input shows great promise for improving current detection and classification of astronomical objects.

2. Defining the Data

To help us identify the SDSS photometric attributes that show promise in merger classification, data from the SDSS survey were collected in two distinct groups, one group chosen as a representative sample of galaxy objects in SDSS, and the other to represent known mergers.

2.1. **Data Sources.** We utilized data strictly from the Galaxy Zoo project and SDSS. Galaxy Zoo was used to obtain SDSS ID's for merger objects, along with an attribute representing the users' confidence in the classification as a merger. All photometric data, merger or non-merger, was obtained from the SDSS.

2.1.1. *Mergers.* The data chosen to represent known merging galaxies were represented by 2,810 of the 3,003 SDSS mergers presented in [3] (i.e., those that had the full set of attributes that we examined).

These objects are known to be involved in mergers and to represent objects with relatively high surface brightness (making human classification possible).

2.1.2. *Non-Mergers.* To build classification models, galaxies assumed to be predominantly non-mergers were also needed as training examples.

As the vast majority of the 100 million SDSS galaxies are not mergers, a representative random sample of SDSS galaxies was chosen for this role.

The sample (initially comprised of 3500 instances) was chosen at random from objects of galaxy type within the SpecPhotoAll view in the SDSS database. This view represents objects that have spectral data associated with them. The spectral data was necessary to obtain object redshift, which was needed to remove distance dependence from the gathered attributes.

Utilizing objects with spectral data also had the ancillary impact of restricting the non-mergers to those with similar surface brightness to the mergers.

2.2. **Data Cleaning and Pre-Processing.** Upon completion of these steps, the sample consisted of 6,310 objects with 76 attributes, including the nominal attribute "merger/non-merger." Considerable pre-processing was necessary to ready the data for use as the training set for classifiers. Some pre-processing steps were necessary for both of the two algorithms utilized. All attributes that did not represent morphological characteristics were removed. For example, the SDSS object ID's, measurement error magnitudes, and attributes representing location or identity, rather than morphology, were among those removed. In Astronomical Catalog missing values occurs for variety of reason from. It is not possible to estimate these values, as these values may be physically meaningful. Therefore instances with placeholder values (in SDSS, "-9999") in any attribute were

removed. Since data were gathered from bright objects, most objects did not require this removal. Distance-dependent attributes were transformed, using redshift, to be distance-independent. A concentration index was also generated, using the ratio of the radii containing 50% and 90% of the Petrosian flux within each galaxy.

2.3. **Attributes.** *Note: Each of the following attributes typically exists for the five SSDS filter wavebands u, g, r, i, z.*

Attribute	Description
$petroMag_{ug}$	Petrosian magnitude colors. A color was calculated for four independent pairs of bands in SDSS (u-g, g-r, r-i, and i-z).
$petroRad_u * z$	Petrosian radius, transformed with redshift to be distance-independent.
$invConIndx_u$	Inverse concentration index. The ratio of the 50% flux Petrosian radius to the 90% flux Petrosian radius.
$isoRowcGrad\_u * z$	Gradient of the isophotal row centroid, transformed with redshift to be distance-independent.
$isoColcGrad_u * z$	Gradient of the isophotal column centroid, transformed with redshift to be distance-independent.
$isoA_u * z$	Isophotal major axis, transformed with redshift to be distance-independent.
$isoB_u * z$	Isophotal minor axis, transformed with redshift to be distance-independent.
$isoAGrad_u * z$	Gradient of the isophotal major axis, transformed with redshift to be distance-independent.
$isoBGrad_u * z$	Gradient of the isophotal minor axis, transformed with redshift to be distance-independent.
$isoPhiGrad_u * z$	Gradient of the isophotal orientation, transformed with redshift to be distance-independent.
$texture_u$	Measurement of surface texture.
$lnLExp_u$	Log-likelihood of exponential profile fit (typical for a spiral galaxy).
$lnLDeV_u$	Log-likelihood of De Vaucouleurs profile fit (typical for an elliptical galaxy).
$fracDev_u$	Fraction of the brightness profile explained by the De Vaucouleurs profile.

3. Machine Learning

3.1. **Decision Trees.** Decision trees are a straightforward machine learning algorithm that produces a classifier with numerical or categorical input, and a single categorical output (the 'class'). Decision trees have several advantages:

- The resulting tree is equivalent to a series of logical 'if-then' statements, and is therefore easy to understand and analyze.
- Missing attribute values can be incorporated into a decision tree, if necessary.
- Easy to implement as a classifier.
- Computationally cheap to 'train' and use in classification.

The most popular decision tree algorithm, C4.5, was published by Ross Quinlan in 1993 [8]. To generate a decision tree, the Weka data mining software suite was utilized. Weka is a robust and mature open source Java implementation of many prominent machine learning algorithms. It also automates many pre-processing tasks, including transformations of parameters and outlier

detection/removal. Weka refers to its C4.5 implementation as J48. This is the routine we used to build a decision tree for classification.

3.1.1. *Decision Trees in Weka.* The Weka J48 algorithm has several arguments. The relevant arguments for our exploration are:

- **binarySplits**: If set to true, the generated tree will be binary. A binary tree is simpler to interpret.
- **confidenceFactor**: The lower this is set, the more pruning that will take place on the tree. More pruning can result in a simpler tree, at the expense of predictive power. However, too little pruning can contribute to overfitting.
- **minNumObj**: The minimum number of instances required in each tree leaf. The higher this is set, the simpler the resulting tree.

As the goal of this work is primarily to explore the strength of SDSS attributes in merger classification, emphasis in tree generation was on generating simple trees, and examining the strongest predicting attributes. In particular, we are searching for those database attributes that contain the most predictive power: those that show the highest correlation with Galaxy Zoo volunteer-provided classification as a merger. These would be the attributes that match most strongly with the outputs of human pattern recognition.

3.1.2. *Information Gain.* In the C4.5 and J48 algorithms, the tree design is predicated upon maximizing information gain (a measurement of entropy in the data). Using Weka, the information gain was calculated for each of the attributes, using the 6310 instances referenced in section 2.2 with tenfold cross-validation. The top five attributes are listed below. Notably, 4 of these top 5 attributes are related to the SDSS observations in the green waveband. These are the attributes that have the highest predictive power in merger classification accuracy.

Attribute	Information Gain
$lnLExp_g$	0.099
$texture_g$	0.074
$lnLDeV_g$	0.068
$petroMag_{gr}$	0.065
$isoAGrad_u * z$	0.057

3.1.3. *Decision Tree Results.* We decided to generate three different trees, with the following characteristics:

(1) A tree that is trained on all instances. This tree should use all mergers, regardless of the vote of merger confidence given by Galaxy Zoo users.
(2) A tree that is trained on merger instances with stronger Galaxy Zoo user confidence. This tree was to be generated with only mergers that a majority of Galaxy Users flagged as such. These instances are assumed to be the mergers that are, in some sense, 'obvious.'
(3) A tree that is trained on merger instances with less than a majority of Galaxy Zoo users indicating then as such. These instances are assumed to be less than obvious to the layperson.

If one simply classifies all galaxies as non-mergers, a predictive accuracy of 55% is obtained. In the simplest tree with one split (seen in figure 1), a 66% correct classification occurs, so there is a modest but definite information gain. The attribute $lnLExp_g$ is at the root node with values at or below -426.586609 indicating a merger and all others classified as non-mergers.

When the minimum number of leaf instances is set to 500, and the confidence factor to 0.001, a relatively simple tree is obtained that still has a reasonable predictive power of 70%. A 66%/34% training/test set split was used. A portion of the model output is shown below.

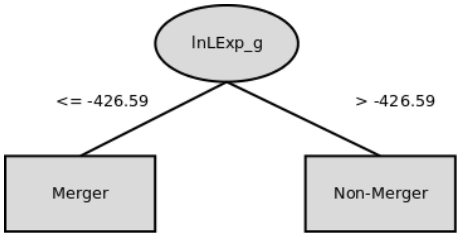

FIGURE 1. Visualization of decision tree with a single node.

	Precision	Recall	F-Measure
Merger	0.659	0.682	0.670
Non-Merger	0.734	0.714	0.724
Weighted Avg.	0.700	0.699	0.700

The root node of this tree (as seen in figure 2) is $lnLExp_g$, which is not a wholly unexpected result, as will be discussed later in this paper.

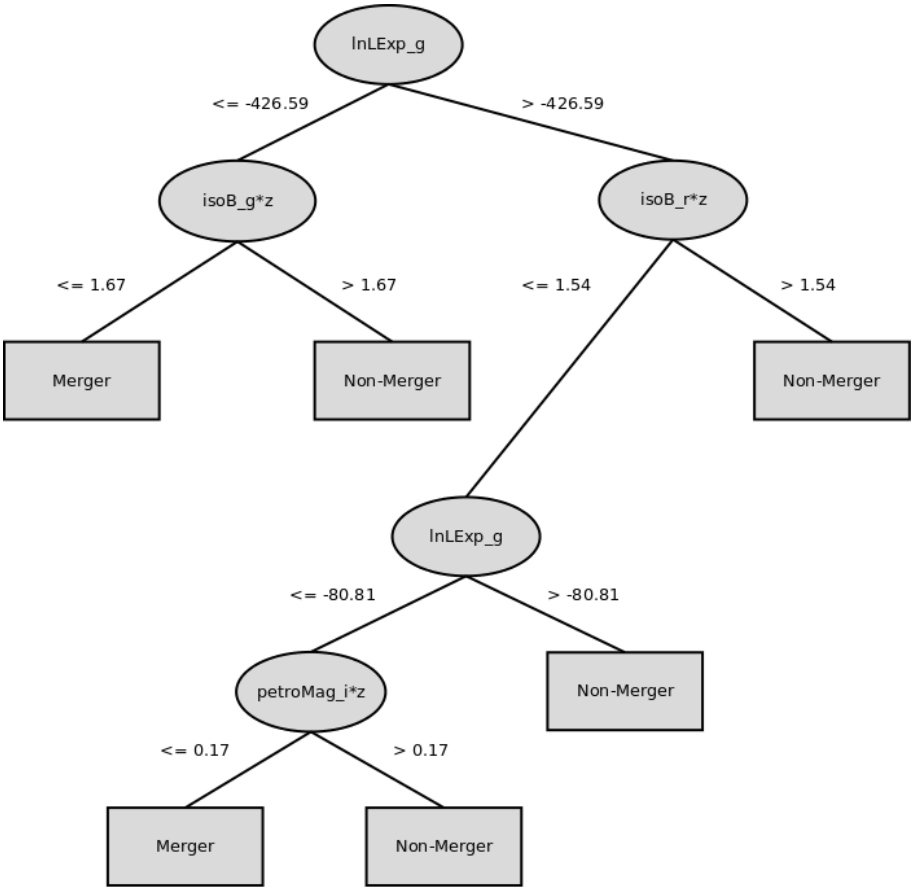

FIGURE 2. Visualization of decision tree built using all mergers.

After removing merger instances with a user confidence of less than 0.50 (with the number of leaf instances set to 200 to produce a simple tree and a 66%/34% split),we measured the precision, recall and F-measure for each of the two classes to determine the accuracy of the model. For mergers,

recall is calculated as the proportion of the number of mergers correctly classified as such out of the total number of mergers. Precision is calculated as the proportion of the number of mergers correctly classified as such out of all instances classified as mergers (correctly or not). The F-measure is a commonly reported measure intended to incorporate both precision and recall into a single measure. It is defined as $\frac{2 \cdot precision \cdot recall}{precision + recall}$.

	Precision	Recall	F-Measure
Merger	0.657	0.456	0.538
Non-Merger	0.766	0.882	0.820
Weighted Avg.	0.730	0.741	0.726

Contrary to intuition, while the overall classification accuracy increases, the recall of the model for mergers decreased significantly. With this approach, $petroMag_{gr}$ is now the strongest predictor at the root of the tree. This can be seen in figure 3. $lnLExp\_g$ is still a key attribute, but it is no longer at the root. This model has very strong predictive power for non-mergers, but quite weak recall for mergers.

FIGURE 3. Visualization of decision tree built using the strongest mergers.

After removing merger instances with a user confidence of more than 0.50 (with the number of leaf instances set to 200 to produce a simple tree and a 66%/34% split), we achieve the output shown below.

	Precision	Recall	F-Measure
Merger	0.416	0.167	0.238
Non-Merger	0.796	0.933	0.859
Weighted Avg.	0.712	0.762	0.721

The users' confusion seems to be expressed in the resulting model, which has high overall accuracy, but a very weak recall. This poor performance is due to its excessive tendency to classify as Non-Merger, as the data set now is only comprised of objects that are not obviously mergers. Using these weaker voted mergers, the model is rooted on $petroMag_{ui}$, as seen in figure 4.

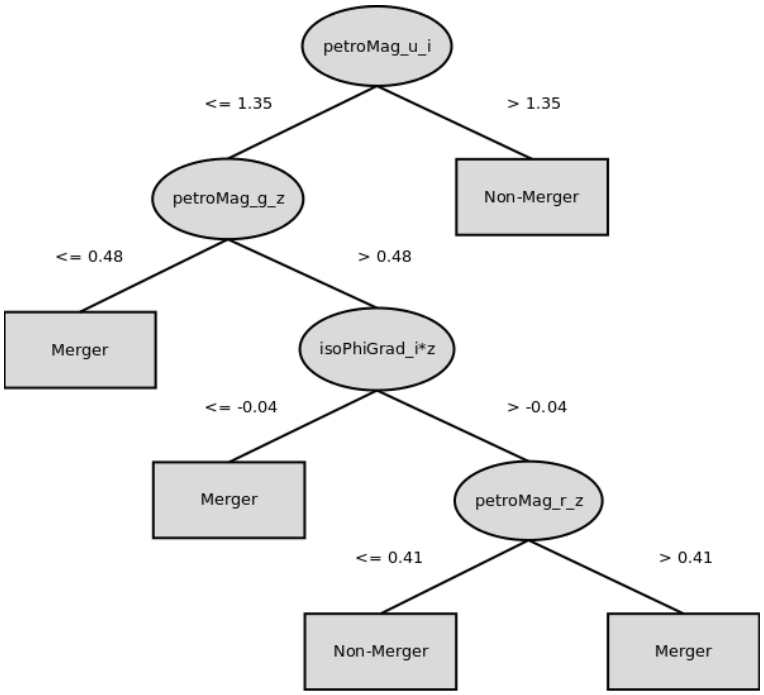

FIGURE 4. Visualization of decision tree built using the weakest mergers.

3.1.4. *Tree Strengths and Weaknesses.* The trees generated are of varying usefulness.

The tree generated using all of the mergers exhibited an overall accuracy of about 70%, with precision of 66% and recall of 68%. This is above average predictive power, but not incredibly useful.

The trees generated using the stronger and weaker mergers separately seem to indicate two things:

(1) The user confusion over some mergers appears to be manifested in the resulting model, as the parameters that are influential in the model are not strongly morphological, indicating that the objects may be missing strong visual cues of merging.
(2) The confidence of users in some merger classifications results in a tree that incorporates more strongly morphological attributes, but has diminished recall power. We feel that this merits further investigation.

There are two especially interesting things about the decision trees generated from this data:

- The strongest predicting attributes seem to be associated with the SDSS green filter wave-band.
- Poor exponential fit and small isophotal minor axis are among the strongest indicators of merger presence.

3.1.5. *Significance of the Green Band.* The strongest predicting attributes seem to be associated with the green band. In the tree generated using all merger instances, The two strongest attributes for merger prediction are associated with the green band, and fully half of the top ten information gaining attributes are associated with this band. The green band seems to carry a disproportionate amount of information relative to the other four bands measured in SDSS photometry.

Upon investigation, we discovered that strong green spectral lines are associated with stellar formation via doubly ionized oxygen, and stellar formation is itself unusually abundant in galactic mergers[7]. So it is not surprising that the green band seems to be important in the classification models we have generated.

3.1.6. *Significance of lnLExp and isoB Attributes.* The attributes $lnLExp$ and $isoB$ both featured prominently in the decision tree approach as influential values for merger detection.

The $isoB$ attribute represents the length of the minor axis of the isophote of the galaxy's surface brightness in a given band. It is a reasonable expectation that tidal distortion from merger involvement may influence an axis of such an isophote.

The $lnLExp$ attribute represents the extent to which the galaxy object has a brightness profile that is fit well by an exponential fit, the details of which can be found in [9]. It is not surprising that this measure of morphology would be an influential factor in merger classification, as tidal distortion would almost certainly affect the brightness profile of a galaxy involved in a merger and thereby reduce the likelihood that the galaxy brightness profile would be fit by a standard non-distorted spiral galaxy exponential function. It should also be noted that another measure of brightness profile fit was featured among attributes with the highest information gain: $lnLDeV$. $lnLDeV$ is a measure of goodness of fit with the De Vaucouleur profile (which is the functional form of the brightness profile in elliptical galaxies), and this would also be expected to exhibit irregularities in the presence of tidal distortion in true colliding/merging galaxies.

3.1.7. *Future Direction for Decision Trees.* Given the modestly strong evidence that we have generated for the quality of green-band morphological attributes as merger predictors, a promising avenue for further development of classifiers may be other attributes in this band. These may be novel image characterization parameters or simply transformations of existing database parameters.

The inclusion of isophotal axis length among the influential parameters seems to indicate that more examination of isophotal properties may be fruitful in this area.

4. Cluster Analysis

Identifying groups of similar observations in a dataset is a fundamental step in any data analysis task. Classification and clustering are the two main approaches used to identify similar groups of data instances. Whereas classification attempts to assign instances to one of several known classes, clustering attempts to derive the classes themselves. In the case of one or two dimensions, visual inspections of the data such as scatter plots can help to quickly and accurately identify the classes. Datasets in astronomy are generally comprised of many more dimensions. With advancements in astronomical data collection technology, astronomers are able to collect several hundred variables for millions of observations. Not all these collected variables are useful for a given classification task. There typically are many insignificant attributes that might prevent us from identifying the structure of the data.

With the knowledge of class labels from the Galaxy Zoo catalog of merging and interacting galaxies, we would like to be able to identify which morphological and photometric attributes in the SDSS data correlate most strongly with the user-selected morphological class. These variables can be identified by measuring the separation of the instances in the attribute feature space in which the data reside: which attributes provide the best discriminator between different human-provided patterns and classes? Measures like Dunn's Validity Index[4] and Davies-Bouldin Validity Index[5] are two metrics by which to achieve this.

4.1. **The Davies-Bouldin Index.** Davies-Bouldin Validity Index (DBI) is a function of the ratio of *intra*-cluster instance separation to *inter*-cluster instance separation. This is given by:

$$DB = \frac{1}{n}\sum_{i=0}^{n} max_{i\neq j}\frac{S_n(Q_i)+S_n(Q_j)}{S(Q_i,Q_j)}$$

...where n is the number of clusters, $S_n(Q_i)$ is the average distance of all objects from the cluster to their cluster center, and $S(Q_i, Q_j)$ is the distance between clusters centers. Good clusters (i.e., compact clusters with respect to their separation) are found with low values of DBI, and poor clusters (i.e., strongly overlapping groupings) have high values of DBI. For the inter-cluster distance function S one could use single linkage, complete linkage, average linkage, centroid linkage, average of centroids linkage, or Hausdorff metrics and for the intra-cluster distance function S one could use complete diameter, average diameter, or centroid diameter[4]. For purposes of experimentation, we picked used the centroid linkage and the centroid diameter as our measures to calculate the DB index.

4.2. **Approach.** To determine the database attributes that influence the separation of the human-provided galaxy classes (merger versus non-merger) most strongly, we first calculated the DB index for the two clusters (i.e., the cluster of mergers versus the cluster of non-mergers) using each one of variables individually. We then ranked the variables based on these calculated DBI values. The variable that tops this list is the most important variable for instance separation, at least according to this metric. This single variable of course cannot necessarily provide us with the best separation. So we looked for any higher dimensional subset of the feature space that has improved separation for these two classes of objects. To this end, we selected the top ten individual variables and calculated the DB index of all possible combinations of these ten variables and ranked the combinations to identify the subset of the original attribute set that provides the best separation.

4.3. **Results.** The following is the list of the top 10 features and subsets with the lowest DB index:

10 Best Separating Individual Attributes	10 Best Separating of all 1014 Subsets of Best 10 Attributes
$isoAGrad_u * z$	$isoAGrad_u * z$
$petroRad_u * z$	$petroRad_u * z$
$texture_u$	$texture_u$
$isoA_z * z$	$isoA_z * z$
$lnLExp_u$	$lnLExp_u$
$lnLExp_g$	$lnLExp_g$
$isoA_u * z$	$petroRad_u * z$, $isoB_z * z$, $isoBGrad_u * z$, $lnLExp_g$
$isoB_z * z$	$isoAGrad_u * z$, $lnLExp_g$
$isoBGrad_u * z$	$petroRad_u * z$, $isoA_u * z$, $isoB_z * z$, $lnLExp_g$
$isoAGrad_z * z$	$isoAGrad_u * z$, $isoBGrad_u * z$, $lnLExp_g$

Features such as $isoPhiGrad_i * z$, $isoColcGrad_g * z$, $isoColcGrad_u * z$, $petroMag_{ug}$, $isoColcGrad_i * z$, and $fracDev_z$ have a significantly large DBI and are therefore do not appear to be useful for clustering. These features seem to be of little significance for decision tree classification as well, since they were not present in any of the trees we generated. Also, visual inspection of the attributes using histograms revealed that with the four individual attributes with lowest DB Index values (seen in figure 5), little to no separation can be seen.

In the scatter plot (seen in figure 6) of mergers and non-mergers in $isoAGrad_u * z$, $lnLExp_g$ feature space shows slight separation between these two classes.

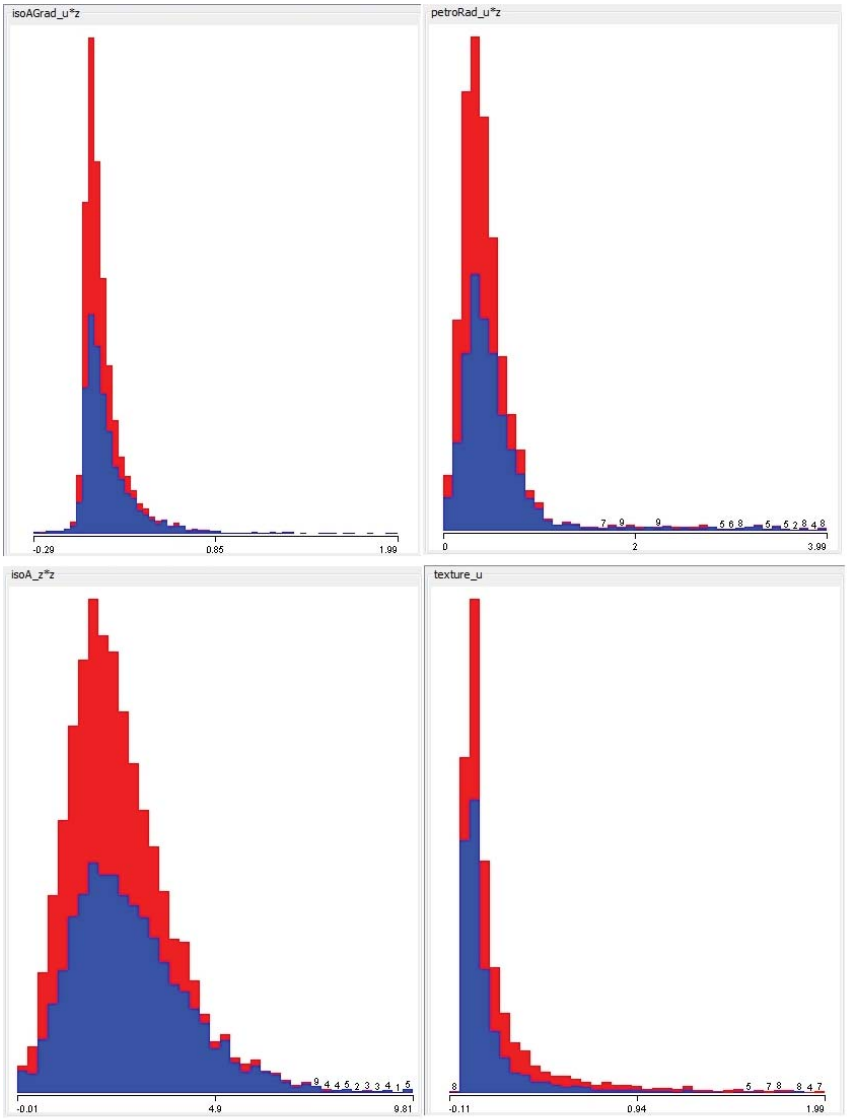

FIGURE 5. Histograms of the four lowest attributes according to DBI.

4.4. **Future Direction for Cluster Analysis.** From the plots it is evident that there is not a clear separation between mergers and non-mergers in the subsets of the feature space that we have explored. This is also evident from the fact that the minimum value of all DBI's that we calculated is 2.19, which is substantially greater than the ideal value of 1. This is an indication of relatively weak clustering. The value 2.19 is the local minimum of the parameter-space. With further analysis of all the possible (75-factorial!) combinations of the 75 numerical attributes, we might be able to find the global minimum value where the clusters have the strongest separation. However, finding the global minimum in this way would be extremely (in fact, prohibitively) computationally intensive. It is, however, important to note that two of the top ten features according to individual DBI are $isoAGrad_u * z$ and $lnLExp_g$, which are also among the top five features in information gain. Therefore, our approach to feature extraction is to some degree consistent with the information

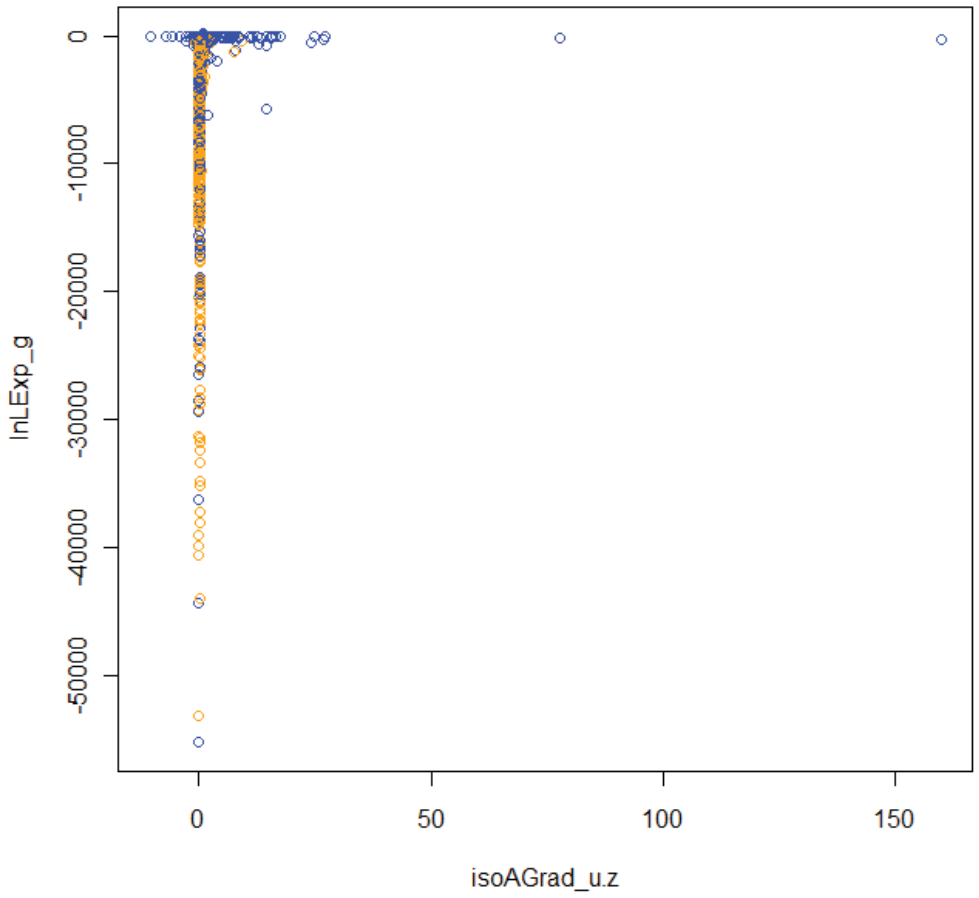

FIGURE 6. Merger and non-merger classes in $isoAGrad_u * z$, $lnLExp_g$ space.

gain-based decision tree approach. With limited computation time and resources, only certain combinations of the best ten attributes could be examined. Use of optimal search algorithms (such as genetic algorithms) and use of a massively parallel computational environment (such as Cloud computing) could empower us to discover the best separating subset of the attributes and provide some interesting results.

5. SUMMARY OF OUTCOMES

We were able to generate a decision tree with accuracy of approximately 70%, including recall for merger detection of approximately 66%. Two classes of morphological attributes were identified as potentially having promise in future work on decision tree analysis:

- Attributes related to the SDSS green waveband, specifically brightness profile fits in this band. This result is validated by the known characteristics of star formation emissions in merging galaxies.
- Attributes related to the galaxy isophotes. This has validity due to the tidal distortions of isophotes that are typically present in galactic mergers.

Results from the cluster analysis also indicate the significance of these two feature-types, providing more evidence of their importance in merger classification. Further analysis might lead to combinations of features that greatly improve the classification accuracy of mergers and non-mergers. Mathematically derived or entirely novel features (especially of a more morphological nature) could also be a promising avenue for improving merger classification, as success with the chosen features was modest. Utilizing a combination of cluster-based feature extraction and decision tree analysis will likely aid in further improvements to classification accuracy, and more generally, to the identification of the salient features that will enable automated pipelines to emulate human cognitive powers and pattern recognition abilities, and thereby automatically indicate the presence of such events in massive petascale sky surveys of the future.

6. Acknowledgements

This research is supported in part by NSF through award #0941610 and in part by NASA through the American Astronomical Society's Small Research Grant Program.

References

[1] N. M. Ball, R. J. Brunner, A. D. Myers, and D. Tcheng. Robust Machine Learning Applied to Astronomical Data Sets. I. Star-Galaxy Classification of the Sloan Digital Sky Survey DR3 Using Decision Trees. *apj*, 650:497–509, Oct. 2006.

[2] M. Banerji, O. Lahav, C. J. Lintott, F. B. Abdalla, K. Schawinski, S. P. Bamford, D. Andreescu, P. Murray, M. J. Raddick, A. Slosar, A. Szalay, D. Thomas, and J. Vandenberg. Galaxy Zoo: reproducing galaxy morphologies via machine learning. *arXiv*, pages 663–+, Apr. 2010.

[3] D. W. Darg, S. Kaviraj, C. J. Lintott, K. Schawinski, M. Sarzi, S. Bamford, J. Silk, R. Proctor, D. Andreescu, P. Murray, R. C. Nichol, M. J. Raddick, A. Slosar, A. S. Szalay, D. Thomas, and J. Vandenberg. Galaxy Zoo: the fraction of merging galaxies in the SDSS and their morphologies. *arXiv*, 401:1043–1056, Jan. 2010.

[4] D. Davies and D. Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1:224–227, 1979.

[5] J. Dunn. Well separated clusters and optimal fuzzy-partitions. *Journal of Cybernetics*, 4:95–104, 1974.

[6] A. Gauci, K. Zarb Adami, and J. Abela. Machine Learning for Galaxy Morphology Classification. *ArXiv e-prints*, May 2010.

[7] I. Strateva, Ž. Ivezić, G. R. Knapp, V. K. Narayanan, M. A. Strauss, J. E. Gunn, R. H. Lupton, D. Schlegel, N. A. Bahcall, J. Brinkmann, R. J. Brunner, T. Budavári, I. Csabai, F. J. Castander, M. Doi, M. Fukugita, Z. Győry, M. Hamabe, G. Hennessy, T. Ichikawa, P. Z. Kunszt, D. Q. Lamb, T. A. McKay, S. Okamura, J. Racusin, M. Sekiguchi, D. P. Schneider, K. Shimasaku, and D. York. Color Separation of Galaxy Types in the Sloan Digital Sky Survey Imaging Data. *The Astronomical Journal*, 122:1861–1874, Oct. 2001.