EPiC Series in Computing

Volume 76, 2021, Pages 29–38

SEDE 2020. 29th International Conference on Software Engineering and Data Engineering

Detection of unknown galaxy types in large databases of galaxy images

Venkat Margapuri[1], Basant Thapa[2] and Lior Shamir[3]

[1] Kansas State University, Manhattan, Kansas, USA
marven@ksu.edu
[2] Kansas State University, Manhattan, Kansas, USA
thapa@ksu.edu
[3] Kansas State University, Manhattan, Kansas, USA
lshamir@ksu.edu

Abstract

Modern digital sky surveys utilize robotic telescopes that collect extremely large multi-PB astronomical databases. While these databases can contain billions of galaxies, most of the galaxies are "regular" galaxies of known galaxy types. However, a small portion of the galaxies is rare "peculiar" galaxies that are not yet known. These unknown galaxies are of paramount scientific interest, but due to the enormous size of astronomical databases they are practically impossible to find without automation. Since these novelty galaxies are, by definition, not known, machine learning models cannot be trained to detect them. In this paper, an unsupervised machine learning method for automatic detection of novelty galaxies in large databases is proposed. The method is based on a large and comprehensive set of numerical image content descriptors weighted by their entropy, and the farthest neighbors are ranked-ordered to handle self-similar peculiar galaxies that are expected in the very large datasets. Experimental results using data from the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS) show that the ability of the method to detect novelty galaxies outperforms other shallow learning methods such as one-class SVM, Local Outlier Factor, and K-Means, and also newer deep learning-based methods such as auto-encoders. The dataset used to evaluate the method is publicly available and can be used as a benchmark to test future algorithms for automatic detection of peculiar galaxies.

1 Introduction

In the past two decades, Earth-based astronomical instruments have largely shifted from manually controlled telescopes to robotic telescopes that survey and image the entire sky [1], making their data available to the astronomy community through virtual observatories [2]. Instead of traveling to a telescope site and pointing the telescope at the target of their interest, an astronomer can easily use pre-collected sky survey data to study any target of their choice. These powerful imaging instruments generate some of the world's largest databases, contain billions of astronomical objects, and lead to numerous scientific discoveries that were not possible in the pre-information era. Sloan Digital Sky Survey (SDSS) alone has produced data leading to more

than $3 \cdot 10^4$ peer-reviewed papers, and it is very reasonable to assume that more discoveries of paramount scientific interest are hidden inside these databases. However, effective mining of these databases requires powerful algorithms that can process these complex data and turn them into knowledge and scientific discoveries.

One of the effective scientific tasks enabled by digital sky surveys is the identification of unknown objects among the billions of "regular" objects in the databases. Most extra-galactic objects belong in the galaxy classification scheme, known as the "Hubble sequence" [3]. However, some galaxies do not fit any stage on the Hubble sequence and are considered "peculiar" galaxies [4]. Although these galaxies are rare, they are of high scientific interest as they carry important information about the past, present, or future universe.

The Panoramic Survey Telescope and Rapid Response System (Pan-STARRS) is an array of two robotic telescopes synchronized to observe the same part of the sky simultaneously to increase the cost-effectiveness of its imaging power. Launched in 2008, Pan-STARRS used its wide 3^o field of view and 1.4 Gigapixel digital camera to image over $3.5 \cdot 10^9$ astronomical objects and generated the world's largest astronomical database of ∼1.6PB.

While Pan-STARRS is an invaluable source of scientific data, the size of the database and the complex nature of the data makes it highly impractical to analyze manually, reinforcing the need for algorithms that can analyze the data and turn them into scientific discoveries.

In this paper the task of identifying novelty astronomical objects automatically is investigated. Deep-learning based auto-encoders technique is compared to statistical methods based on "shallow learning",

The paper proposes a novelty detection algorithm that uses the concept of entropy of a set of pre-defined numerical image content descriptors. The performance of the algorithm is compared against the performance of common "traditional" unsupervised machine learning algorithms such as One-Class Support Vector Machines (OCSVM), K-Means Clustering, Local Outlier Factor (LOF), and K-Nearest Neighbors algorithm which falls in the realm of supervised learning. In addition, the deep learning technique of auto-encoders is applied and investigated.

2 Related work

Relevant research in the area of study, while not abundant, is existent and studied to help pave a segue for the current work. The first attempt to identify peculiar galaxies on data from the Sloan Digital Sky Survey (SDSS), a sky survey with data analysis, faces challenges that largely overlap with the data analysis challenges of Pan-STARRS. It was done by using a large number of "citizen scientists" who observed the images manually over several years and determined whether the astronomical object is peculiar [5]. That initiative allowed the compilation of a large catalog of rare ring galaxies [6]. However, statistical analysis using ring galaxies detected automatically showed that many more ring galaxies were hidden inside [7]. Additionally, after several years of work involving over 10^5 volunteers, less than 10^6 objects were observed [8]. Applying the same method for the analysis of all objects imaged so far by Pan-STARRS will require over $\sim 10^4$ years to complete.

The size of the data of digital sky surveys reinforces the use of automation. An example of automatic outlier detection applied to datasets of astronomical objects is the application of outlier detection to SDSS galaxy data to identify galaxies with unusual spectroscopic profile [9]. The method is based on unsupervised Random Forest [10], and was applied on the spectroscopic data of the galaxies rather than their images.

Substantial research has been done for general outlier detection. Among numerous approaches, the concept of entropy of features was used to mine outliers in databases [11]. Among

more recent approaches, deep neural networks were used for automatic detection of outliers in data, including image data [12, 13]. While deep artificial neural networks, and in particular deep convolutional neural networks, have shown excellent performance in supervised learning of image data, the use of auto-encoders [12, 13] allows using the power of deep neural networks also for unsupervised machine learning.

3 Data

In the absence of a benchmark with ground truth for novelty galaxy detection, a controlled benchmark dataset of galaxy images from the Pan-STARRS sky survey is compiled. Each image is a 120×120 image in the JPG image format. The benchmark includes three datasets, such that each dataset contains 200 celestial objects. The first contains spiral galaxies, the second contains lenticular galaxies, and the third contains stars. The reason for using stars is that data analysis pipelines of digital sky surveys such as Pan-STARRS often struggle to classify between stars and galaxies, and therefore more objects identified as galaxies are in fact stars. Therefore, a practical algorithm for novelty galaxy detection needs to handle the existence of stars identified incorrectly as galaxies.

The datasets are used such that in each run 200 galaxies from one dataset are combined with 10 galaxies from another dataset to create a dataset in which the majority of the galaxies are "regular" galaxies, but a small number of galaxies which are different from the majority of the galaxies are also included. That allows to develop and test methods for identifying galaxies that are different from most other galaxies. For instance, in a late-type universe that contains only spiral galaxies, a lenticular galaxy would be considered a rare novelty galaxy. Similarly, in a universe of just stars, a lenticular galaxy is considered peculiar. Therefore, it can be reasonably assumed that an unsupervised machine learning algorithm that is not trained on spiral galaxies yet automatically detects a small number of spiral galaxies among a large number of lenticular galaxies, is an algorithm that will also be able to identify other novelty galaxies without training. Figure 1 shows examples of the celestial objects as imaged by Pan-STARRS.

Figure 1: Example image of star (left), lenticular galaxy (center) and spiral galaxy (right) imaged by Pan-STARRS

The dataset is freely available at PanSTARRSData, and can be used as a benchmark dataset for developing future algorithms for automatic detection of novelty galaxies.

4 Method

According to shallow supervised learning of image data, each image in the dataset is first converted to a set of numerical image content descriptors that reflect its visual content through numerical values. The set of numerical image content descriptors used in this study is WND-CHARM [14], that was proven effective to machine analysis of galaxy images [15, 16, 17, 18]. In summary, the WND-CHARM library computes a comprehensive set of 2883 numerical image

content descriptors that reflect numerous aspects of the visual content such as the shape, color, edges, textures (e.g., Gabor, Haralick, Tamura), fractals, polynomial decomposition of the image (e.g., Chebyshev polynomials, Zernike), and statistics of the distribution of the pixel values (e.g., Radon features, multi-scale histograms, first four moments). That feature set is described in detail in [14, 19, 20], and is applied successfully to the task of galaxy image analysis [21, 22].

The feature extraction process computes 2883 numerical image content descriptors for each galaxy image. That large set is sufficiently comprehensive to reflect numerous aspects of the galaxy morphology [15, 16, 17, 18]. However, it can also be assumed that many of these descriptors are not informative for unsupervised detection of novelty galaxies, and possibly add noise to the system.

In order to select the most informative features and avoid noise to better detect novel objects in the dataset, a process of feature selection is required. Since the learning is unsupervised, many "traditional" feature selection algorithms are not suitable. Therefore, in this study, the concept of Entropy is used as a technique to perform unsupervised feature selection on datasets with a large number of features.

The entropy of a system S with N possible outcomes is computed as - $\sum_{i=1}^{N} p_{\mathrm{i}} \cdot log(p_{\mathrm{i}})$, where p_i is the frequency of outcome i in S. To compute entropy on the numerical image content descriptors, the value of each numerical content descriptor is convolved into a histogram of N bins, and p_i is the frequency of the values in the histogram bin i, such that $i \in \{1..N\}$. The intuition behind this method of feature selection is that informative features tend to have their values distributed in some non-random clusters of values, while non-informative features have their values randomly distributed.

Identification of novelty galaxies is unique in the sense that due to the enormous size of the datasets of galaxy images, a single one-of-a-kind peculiar galaxy is unlikely to exist. For instance, the future Vera Rubin observatory is expected to collect $\sim 10^{10}$ galaxies, and therefore even an extremely rare one-in-a-million object is expected to appear in the dataset about 10^4 times. Therefore, an effective novelty galaxy detection algorithm is required to be sensitive to the number of galaxies in the dataset, and assume that many of the novelty galaxies are self-similar to each other.

To handle the self-similarity of novelty galaxies, the intuition of the algorithm is that, given a set of galaxies, the farthest K^{th} neighbor amongst the K^{th} nearest neighbors of all the galaxies is a novelty galaxy. This allows the user of the algorithm to specify a minimum number of self-similar novelty galaxies. For example, consider a dataset of 100 galaxies with a K value of 10. The distance of each galaxy in the dataset is determined by its 10^{th} nearest neighbor. Therefore, if a galaxy has nine similar neighbors but is different from the remaining 90 galaxies, it will be assigned a high distance that reflects its dissimilarity from most of the galaxies. This simple mechanism might be inferior to other algorithms for the general case of novelty detection, but it is suitable for the detection of novelty galaxies as it provides the user with clear control over the number of self-similar novelty galaxies. This number changes with the type of galaxies considered, and therefore, the user is required to adjust the number based on the size of the dataset and the estimated frequency of different types of novelty galaxies.

The algorithm is described as follows:

1. Normalize the values in the dataset using Min-Max normalization.
2. Compute the entropy of each of the features of the dataset.
3. Choose a value between 0 and the greatest entropy of the features as the entropy Threshold.
4. Apply the entropy threshold to the entropies of the features and set all entropies greater than the threshold to 0.

5. Pick a K, the order of the neighbor to be considered as the nearest neighbor. For instance, if the value of K is set to 5, the distance to the 5^{th} closest neighbor of each of the galaxies is used as the dissimilarity measure of that galaxy.
6. Compute the distance to the K^{th} neighbor of each of the galaxies using Minkowski distance i.e. weighted Euclidean Distance where the weights of the features are the entropy values obtained in Step 4.
7. Sort the galaxies by their distance to their K^{th} neighbor. Greater the distance, higher the likelihood that the galaxy is a novelty.

The algorithm depends on two parameters that control its performance:
1. The order of the closest neighbor (K): If the value of K is lower than the number of novelty galaxies of a specific type, it is possible that the distance between a certain galaxy and its K^{th} neighbor is not larger than other non-novelty galaxies. Therefore, the user is required to select a value that is higher than the number of novelty galaxies of a certain type that are expected to exist in the dataset. The number depends on the size of the entire dataset and also not necessarily known to the user. In that case the user will need to attempt several K values and inspect the results to see if the detected novelty galaxies are indeed not "regular" galaxies.
2. The value of the entropy threshold (Step 3 in the algorithm above): A high entropy threshold might lead to the rejection of features that carry information about the morphology of the galaxy. On the other hand, a low threshold might lead to the inclusion of noisy features.

The source code of the algorithm can be found at PanSTARRSNoveltyDetectionAlgorithm.

5 Results

To test the performance of the method, the algorithm described in Section 4 is applied to the data described in Section 3. Figure 2 shows the performance of the algorithm when the K parameter discussed in Section 4 is set to 5, 10, and 20. The performance of the method is tested on different ranks.

The rank r is the number of query galaxies determined by the algorithm as the most likely to be novelty galaxies. If a novelty galaxy is among these r galaxies, the attempt is considered a hit, and otherwise a miss. Since candidates of novelty galaxies are inspected manually, a method that returns false positives is acceptable as long as the novelty galaxies are among a set that is small enough for manual analysis. Note that the problem of novelty galaxy detection does not require identifying all novelty galaxies, as novelty galaxies of the same type are expected to be present multiple times in galaxy datasets acquired by robotic telescopes.

The results show that the performance of the algorithm when identifying spiral galaxies among lenticular galaxies is better than the performance of the algorithm when identifying stars among lenticular galaxies. This is partly explained by the fact that lenticular galaxies and stars are more similar in morphology to each other compared to lenticular and spiral galaxies.

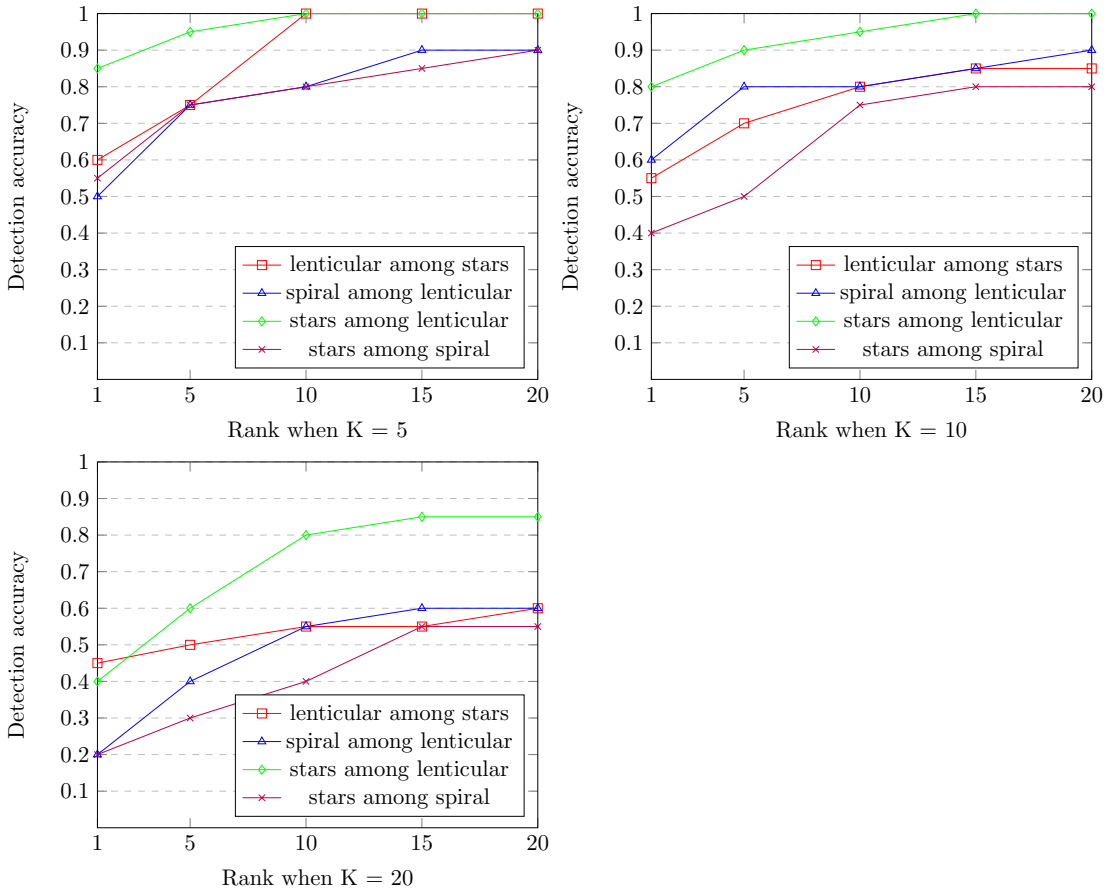

Figure 2: Detection accuracy when using different datasets and ranks.

6 Comparison to novelty detection algorithms

Since the problem of automatic novelty galaxy detection is relatively new, not many proposed novelty detection algorithms for galaxies are available in the existing literature. Hence, the performance of the proposed algorithm is compared against "traditional" novelty detection algorithms such as one-class SVM, K-Means, and Local Outlier Factor (LOF), as well as the deep learning-based auto-encoders.

6.1 Comparison to deep learning with auto-encoders

Auto-encoders [12] are a class of unsupervised machine learning using artificial neural networks (ANN). A typical artificial neural network consists of an input layer, which inputs the data to the layers of the neural network, several hidden layers, and an output layer, which outputs the outcome. Each of the hidden layers in the network performs computations on the weighted inputs and transfers the computed result to the next layer. An auto-encoder can be conceptualized as a specific type of neural network that copies the input values to the output without requiring a target variable. Since target variables are not required, it is a good fit for

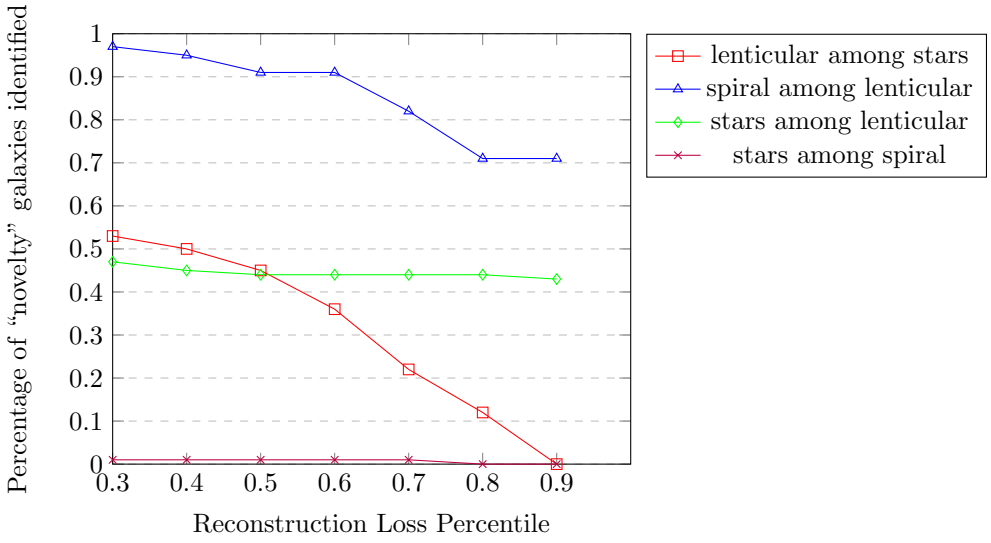

Figure 3: Performance of novelty galaxy detection using the autoencoder reconstruction loss

unsupervised learning [13].

For this experiment, a deep auto-encoder is used. The auto-encoder architecture comprises of ReLU activation function in the encoding layers and sigmoid activation function in the decoding layers. The loss function used is binary cross-entropy and the optimizer used is RMSProp. The size of the input of 120×120,

Auto-encoders with two different architectures are developed. One, with hidden layers of sizes 128, 64, 32, 64, 128, and the other with hidden layers of sizes 1024, 512, 256, 512, 1024.

In each of the datasets, the "regular" galaxy images are split into two groups, one containing 180 images to train the auto-encoder, and another of 20 images to test on the auto-encoder to obtain the reconstruction losses. Then, the "novelty" galaxy images are tested on the auto-encoder, and the loss of the "novelty" galaxies is compared to the loss of the "regular" galaxies. For evaluation, the 30^{th} to the 90^{th} percentile of reconstruction loss values on "regular" galaxies are used as thresholds, and the percentage of "novelty" galaxies identified from amongst 200 images of "novelty" galaxies is computed as shown in Figure 3

As the figure shows, the majority of the "novelty" galaxies have reconstruction losses similar to those of the "regular" galaxies. As a result, the auto-encoder model does not necessarily provide a reconstruction loss that distinctly identifies a "novelty" galaxy.

6.2 One-class Support Vector Machines (OCSVM)

The OCSVM algorithm is applied to each of the four datasets using the scikit-learn library.

The performance of the algorithm is measured as the number of actual "novelty" galaxies identified by the algorithm divided by the total number of "novelty" galaxies attempted. Ideally, only the ten "novelty" galaxies are identified as "novelty" galaxies by the algorithm, in which case the detection rate would be 100%. However, the observation on all four datasets is that the algorithm identifies a large portion of "regular" galaxies also identified as "novelty" galaxies while also misidentifying some "novelty" galaxies as regular galaxies. So, the performance of

the algorithm is similar to that of novelty galaxy detection by random chance. The outcomes of the algorithm are shown in Figure 4.

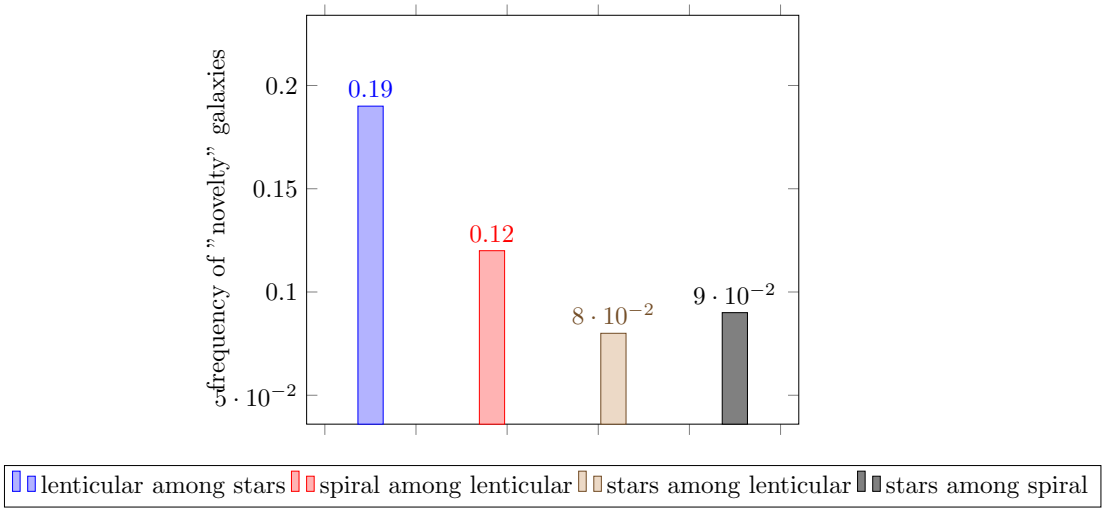

Figure 4: Novelty galaxy detection rate using one-class support vector machine (OCSVM).

6.3 Local Outlier Factor (LOF) algorithm

The Local Outlier Factor (LOF) algorithm produces a score that provides an insight into the likelihood of a data point being an outlier in a given dataset. The scikit-learn LOF library is used to apply the algorithm to each of the datasets. Since the algorithm is unsupervised, no alteration is made to the datasets. A score close to 1 means that the sample is an inlier, while outliers have a larger LOF score. The results show that for each of the datasets, all of the values obtained for the LOF scores are 1, indicating that the algorithm considers all of the images, including the outliers, as the same class as the inliers. As a result, the accuracy obtained using the algorithm is 0 % on all four of the datasets.

6.4 K-Means clustering

K-Means is a simple and established unsupervised learning algorithm which works by choosing a centroid value for each randomly chosen cluster, and iteratively assigning each data point to a cluster that best fits based on the euclidean distance between the data point and the centroids of the clusters. K-Means is typically used for automatic clustering. However, in some cases it can be used for novelty detection by identifying small clusters. If a small cluster is identified, the cluster may contain a small number of self-similar samples that are different from the other samples in the dataset. Therefore, K-Means is an algorithm that could be possibly used for novelty detection in the current scenario. The algorithm is tested with two through 10 clusters. The performance is measured as the number of novelty galaxies among regular galaxies in the cluster in which novelty galaxies are the most frequent. The results are as shown in Figure 5.

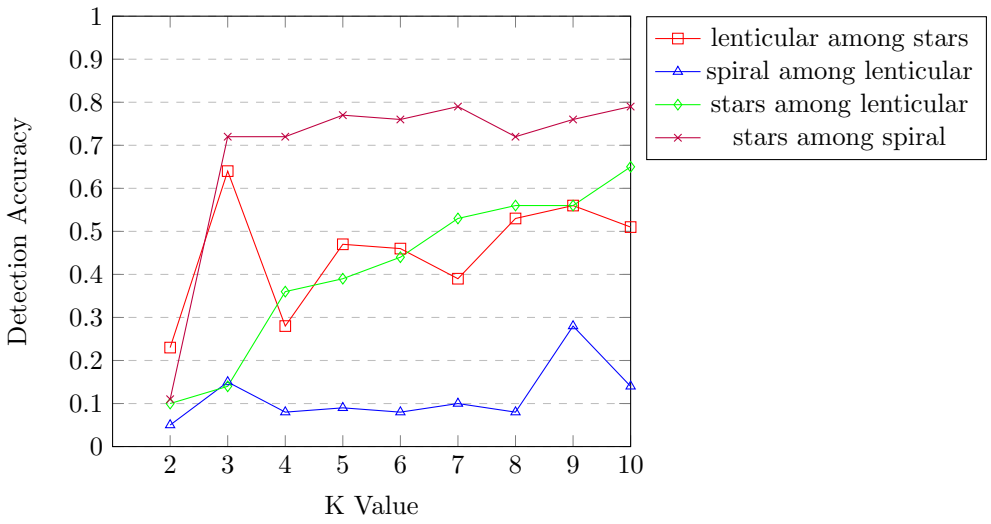

Figure 5: The performance of novelty galaxy detection using the K-Means algorithm

7 Conclusion

Modern robotic telescopes generate some of the world's largest databases. The large size of these databases and a large number of celestial objects makes them impractical to be inspected manually, reinforcing the need for automatic methods to analyze the databases and utilize the important and expensive scientific ventures.

The proposed novelty detection algorithm uses a comprehensive set of numerical image content descriptors, and therefore depends on feature selection. Entropy is shown as a useful way to select features for the problem of unsupervised detection of novelty galaxies.

The method proposed in the paper outperforms "traditional" methods such as one-class SVM, K-Means, and newer methods based on deep neural networks such as auto-encoders. It should be noted, however, that the relatively low number of annotated samples does not allow efficient training of an autoencoder, that normally requires a high number of samples. The dataset used for the experiments is publicly available and can be used for the development and testing of new algorithms for novelty galaxy detection in large astronomical databases.

The downside of the evaluation is that it is performed on a relatively small and controlled dataset, far smaller than the huge datasets generated by modern digital sky surveys. The efficacy of the method will be tested in the future by applying it to extremely large image databases and evaluating its ability to identify real novelty galaxies hidden among millions of celestial objects that have not been inspected yet.

8 Acknowledgement

The research was funded in part by NSF grant number AST-1903823.

References

[1] Kirk Borne. Virtual observatories, data mining, and astroinformatics. Planets, Stars and Stellar Systems, 2013. Springer.

[2] S George, Djorgovski, Mahabal, Ashish, Drake, Andrew, Graham, Matthew, Donalek, and Ciro. Sky surveys. Springer, 2013.

[3] William W Morgan and NU Mayall. A spectral classification of galaxies, 1957.

[4] Arp and Halton. Atlas of peculiar galaxies, 1966.

[5] Chris J Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M Jordan Raddick, Robert C Nichol, Alex Szalay, Dan Andreescu, et al. Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey, 2008.

[6] Ronald J Buta. Galactic rings revisited–i. cvrhs classifications of 3962 ringed galaxies from the galaxy zoo 2 database, 2017.

[7] Lior Shamir. Automatic detection of full ring galaxy candidates in sdss, 2020.

[8] Chris Lintott, Kevin Schawinski, Steven Bamford, Anže Slosar, Kate Land, Daniel Thomas, Edd Edmondson, Karen Masters, Robert C Nichol, M Jordan Raddick, et al. Galaxy zoo 1: data release of morphological classifications for nearly 900 000 galaxies, 2011.

[9] Dalya Baron and Dovi Poznanski. The weirdest sdss galaxies: results from an outlier detection algorithm, 2017.

[10] Tao Shi and Steve Horvath. Unsupervised learning with random forest predictors, 2006.

[11] Ming-Jian Zhou and Jun-Cai Tao. An outlier mining algorithm based on attribute entropy, 2011.

[12] T. Amarbayasgalan, B. Jargalsaikhan, and K. H. Ryu. Unsupervised novelty detection using deep autoencoders and density based clustering, 2018.

[13] Kai Sun, Jiangshe Zhang, Chunxia Zhang, and Junying Hu. Generalized extreme learning machine autoencoder and a new deep neural network, 2017.

[14] Lior Shamir, Nikita Orlov, D Mark Eckley, Tomasz Macura, Josiah Johnston, and Ilya G Goldberg. Wndchrm–an open source utility for biological image analysis, 2008.

[15] Evan Kuminski, Joe George, John Wallin, and Lior Shamir. Combining human and machine learning for morphological analysis of galaxy images, 2014.

[16] Lior Shamir. Automatic morphological classification of galaxy images, 2009.

[17] Lior Shamir, Anthony Holincheck, and John Wallin. Automatic quantitative morphological analysis of interacting galaxies, 2013.

[18] Andrew Schutter and Lior Shamir. Galaxy morphology—an unsupervised machine learning approach, 2015.

[19] Lior Shamir, Tomasz Macura, Nikita Orlov, D Mark Eckley, and Ilya G Goldberg. Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art, 2010.

[20] Lior Shamir. Evaluation of face datasets as tools for assessing the performance of face recognition methods, 2008.

[21] Lior Shamir and John Wallin. Automatic detection and quantitative assessment of peculiar galaxy pairs in sloan digital sky survey, 2014.

[22] Lior Shamir. Morphology-based query for galaxy image databases, 2016.