2022 IEEE 18th International Conference on e-Science (e-Science) | 978-1-6654-6124-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ESCIENCE55777.2022.00061

Neural Network Bias in Analysis of Galaxy Photometry Data

1st Hunter Goddard
Kansas State University, Manhattan, KS, USA

2nd Lior Shamir
Kansas State University, Manhattan, KS, USA

Abstract—Advancements in the ability to collect, store, and access astronomical data have made a major impact on astronomy research in the past two decades. These changes reinforced the need for methodology that can analyze large and complex astronomical databases. One of the common tools used to approach that task is machine learning (ML), and specifically artificial neural networks (ANN). One of the primary downsides of ANNs is that they follow complex and non-intuitive data-driven rules, making it virtually impossible to formally describe the way they analyze the data. Here we analyze possible systematic biases that can affect the results observed when applying ANNs to photometry data. The results show that ANNs can lead to systematic biases in the annotation of the data. These biases are difficult to detect and profile, and can behave in a non-intuitive manner. Therefore, catalogs and data products annotated by neural networks should be used with extra caution.

Index Terms—ML bias, artificial neural networks, photometry, galaxy morphology, sky surveys

I. Introduction

The information era has made a revolutionary impact on astronomy research. For instance, autonomous digital sky surveys have enabled the collection of very large astronomical databases, enabling unprecedented discovery power [1], and that trend is bound to continue [2].

One of the primary outcomes of the data collected by digital sky surveys is photometry. Photometry data for each astronomical object includes measurements such as its color, brightness, size, and more.

Due to the large amounts of data collected by digital sky surveys, manual analysis becomes impractical. Given the complexity and high-dimensionality of the data, one of the common ways to approach the analysis of astronomical data is ML. By using existing ML algorithms, researchers can annotate merely a small part of the data, and apply the algorithms to analyze large datasets by allowing the ML algorithm to extract complex rules driven by the data it was trained with. The purpose of this study is to test possible biases when using neural network analysis of photometry data.

II. Data

We use photometry data from the Sloan Digital Sky Survey (SDSS) for objects that were identified as galaxies. These galaxies were separated into spiral and elliptical galaxies [3]. The SDSS records photometric measurements for the u, g, r, i, and z bands. Color features can be obtained by taking the difference of values in adjacent bands (subtracting the longer wavelength from the shorter one), such as $g - r$ and $r - i$. We compute these color features for the exponential, de Vaucouleurs, and Petrosian profiles to use as inputs for our neural networks. In addition, we include the r-band magnitude and radius for each profile, as well as the radii containing 50% and 90% of the Petrosian flux. Our SDSS data correspond to entries in the *Galaxy* view of data release 17 with the clean photometry flag set. We use 247,427 galaxies, of which 126,110 are labeled as elliptical and 121,317 as spiral.

III. Methodology

Assuming that the Universe is isotropic, the distribution of spiral and elliptical galaxies is expected to be the same regardless of the direction of observation. Here we test whether uneven distribution of the training set in the sky can lead to a bias, reflected by differences in the distribution of the network's elliptical and spiral annotations.

The neural network architecture is determined by comparing models with one to four layers, with each layer having either 16, 32, 48, or 64 artificial neurons, creating 340 different candidate architectures with varying width and depth. Each layer uses rectified linear unit (ReLU) activation function and is strongly regularized with dropout at a rate of 20% in the first layer, and 50% in subsequent hidden layers. At the end of every model is a 2-neuron softmax output layer, corresponding to the two morphological classes.

After selecting a network architecture, we analyze whether the distribution of the training samples in the sky affects the accuracy and distribution of predicted class labels. The dataset is divided into three regions of the sky - the constellations Virgo, Hercules, and Cetus. Neural networks are trained and tested with data from different combinations of these regions.

The number of galaxies of each morphological type within the selected areas is described in Table I. To avoid over-representing any particular region, we select 6,000 galaxies of each type within each area.

TABLE I: Breakdown of Samples by Region and Class

	Virgo	Hercules	Cetus
Elliptical	8,527	7,463	6,759
Spiral	8,295	6,675	7,016
Total	16,822	14,138	13,775

The predictions of the trained models are compared using the binomial test with the null hypothesis that a change in training set location should not significantly alter the distribution of the model's annotations. That assumption is based on

978-1-6654-6124-5/22/$31.00 ©2022 IEEE
DOI 10.1109/eScience55777.2022.00061

the fact that the data are taken from the same survey and that the annotations are taken from the same catalog, and since the Universe is isotropic the location of the galaxies in the training set should not impact the annotations.

IV. Results

The network architecture selection process resulted in a four-layer model with 64, 48, 32, and 32 artificial neurons, respectively, containing a total of 7,154 trainable parameters. We chose to use 2,000 samples (1,000 of each morphological class) for training, allowing us to reserve a total of 10,000 samples per region for model evaluation and comparison.

To determine if the area of the sky that the training data are selected from induces bias in this model, we train neural networks with the same architecture from scratch using different combinations of training data from each constellation. For example, one model is trained with data taken only from the area of Hercules (Table II), while another is trained with elliptical galaxies from Virgo and spiral galaxies from Hercules (Table III), and so on. The resulting nine models are then evaluated by predicting class labels for the reserved test set from each region. These models all obtained a test-set accuracy between 96% and 99%.

TABLE II: Hercules Test Set Confusion Matrix by the Model Trained with Hercules Ellipticals and Spirals

	Elliptical	Spiral
Elliptical	4929	71
Spiral	147	4853

TABLE III: Hercules Test Set Confusion Matrix by the Model Trained with Virgo Ellipticals and Hercules Spirals

	Elliptical	Spiral
Elliptical	4798	202
Spiral	48	4952

Table IV lists the two-tailed p-values from applying the binomial test to each pair of homogeneous annotations (i.e. when the training set region is the same as the test set region) and non-homogeneous annotations for the same test set. Statistically significant values ($p < 0.05$) are highlighted in bold. Although most of these values suggest an insignificant variance, others suggest that some parts of the Universe have a different distribution of elliptical and spiral galaxies. These differences are driven by a bias in the neural network, and do not reflect the real sky. However, a researcher using a catalog that was annotated by using a neural network might not be aware of such bias, as the bias is unexpected and not intuitive.

V. Conclusion

The information era has changed astronomy research by enabling data-driven research enabled by the analysis of very large astronomical databases. Instruments generating vast astronomical pipelines reinforce the need for automatic methods that can annotate the data.

While neural networks are very common in modern astronomy, they should also be analyzed for potential downsides. Here we analyze the potential bias driven by the source of the training samples. Experimental results show that despite using a training set from the same survey, the distribution of the locations of the training samples in the sky affects the annotations. To notice these subtle yet statistically significant biases, the user of the data product needs to be familiar with the specifics of how the model was trained. Therefore, a user of a data product annotated using a neural network might not be aware of such bias, and might therefore reach conclusions that are driven by the bias rather than the real sky.

Biases in data collected by digital sky surveys is a known characteristic of these powerful instruments. For instance, extinction and limiting magnitude are not constant in all directions. The use of neural networks adds an additional bias that is not yet fully known, and can make such biases more complex and more difficult to identify and quantify.

While neural networks provide a useful solution to the annotation of very large astronomical databases, they also have several downsides. Since biases are difficult to identify and profile, data products prepared using neural networks alone should be used with caution.

References

[1] K. J. Edwards and M. M. Gaber, "Astronomy and big data," *Studies in Big Data. Springer*, 2014.

[2] J. A. Tyson, "Cosmology data analysis challenges and opportunities in the lsst sky survey," in *Journal of Physics: Conference Series*, vol. 1290, no. 1. IOP Publishing, 2019, p. 012001.

[3] E. Kuminski and L. Shamir, "A computer-generated visual morphology catalog of 3,000,000 sdss galaxies," *The Astrophysical Journal Supplement Series*, vol. 223, no. 2, p. 20, 2016.

TABLE IV: SDSS Binomial Test P-Values

		Evaluation Region					
Training Region		Virgo		Hercules		Cetus	
Ellipticals	Spirals	Elliptical	Spiral	Elliptical	Spiral	Elliptical	Spiral
Virgo	Virgo	—	—	$9.1 \cdot 10^{-2}$	$1.93 \cdot 10^{-1}$	$5.75 \cdot 10^{-1}$	$3.03 \cdot 10^{-1}$
Virgo	Hercules	$7.11 \cdot 10^{-1}$	$4.96 \cdot 10^{-1}$	**$8.79 \cdot 10^{-3}$**	**$4.87 \cdot 10^{-2}$**	$9.52 \cdot 10^{-1}$	$6.97 \cdot 10^{-1}$
Virgo	Cetus	$4.01 \cdot 10^{-1}$	$2.42 \cdot 10^{-1}$	$4.29 \cdot 10^{-1}$	$6.89 \cdot 10^{-1}$	$1.87 \cdot 10^{-1}$	**$2.78 \cdot 10^{-2}$**
Hercules	Virgo	$6.89 \cdot 10^{-1}$	$7.34 \cdot 10^{-1}$	$2.5 \cdot 10^{-1}$	$2.38 \cdot 10^{-1}$	$3.9 \cdot 10^{-1}$	$3.95 \cdot 10^{-1}$
Hercules	Hercules	$1.8 \cdot 10^{-1}$	$8.18 \cdot 10^{-2}$	—	—	$1.19 \cdot 10^{-1}$	**$3.73 \cdot 10^{-3}$**
Hercules	Cetus	$2.71 \cdot 10^{-1}$	**$1.04 \cdot 10^{-2}$**	$6.53 \cdot 10^{-1}$	$3.32 \cdot 10^{-1}$	$2.0 \cdot 10^{-1}$	**$1.37 \cdot 10^{-3}$**
Cetus	Virgo	$7.79 \cdot 10^{-1}$	$7.19 \cdot 10^{-1}$	**$3.94 \cdot 10^{-2}$**	$1.8 \cdot 10^{-1}$	$5.22 \cdot 10^{-1}$	$3.42 \cdot 10^{-1}$
Cetus	Hercules	$4.18 \cdot 10^{-1}$	$4.71 \cdot 10^{-1}$	**$7.24 \cdot 10^{-4}$**	$6.14 \cdot 10^{-2}$	$8.49 \cdot 10^{-1}$	$4.9 \cdot 10^{-1}$
Cetus	Cetus	$6.67 \cdot 10^{-1}$	$1.56 \cdot 10^{-1}$	**$5.81 \cdot 10^{-4}$**	**$2.99 \cdot 10^{-2}$**	—	—