Similarity Search of Low Surface Brightness Galaxies in Large Astronomical Catalogs

Marcos Tidball
Instituto de Física, Federal University of Rio Grande do Sul
Av. Bento Gonçalves 9500, Porto Alegre, Brazil
marcostidball@gmail.com

Cristina Furlanetto
Instituto de Física, Federal University of Rio Grande do Sul
Av. Bento Gonçalves 9500, Porto Alegre, Brazil

Abstract

Low Surface Brightness Galaxies (LSBGs) constitute an important segment of the galaxy population, however, due to their diffuse nature, their search is challenging. The detection of LSBGs is usually done with a combination of parametric methods and visual inspection, which becomes unfeasible for future astronomical surveys that will collect petabytes of data. Thus, in this work we explore the usage of Locality-Sensitive Hashing for the approximate similarity search of LSBGs in large astronomical catalogs. We use 11670190 objects from the Dark Energy Survey Y3 Gold coadd catalog to create an approximate k nearest neighbors model based on the properties of the objects, developing a tool able to find new LSBG candidates while using only one known LSBG. From just one labeled example we are able to find various known LSBGs and many objects visually similar to LSBGs but not yet catalogued. Also, due to the generality of similarity search models, we are able to search for and recover other rare astronomical objects without the need of retraining or generating a large sample. Our code is available on `https://github.com/zysymu/lsh-astro`.

1 Introduction

Low Surface Brightness Galaxies (LSBGs) are typically dwarf galaxies, but have a much lower density of stars, making them very difficult to detect in astronomical images. They constitute an important segment of the Universe's galaxy population and a significant part of the matter in the local Universe is thought to be "hidden" in such diffuse objects. In the last few years, there has been a revival and growing interest in the field of LSBGs thanks to the discovery of a population of large LSBGs termed ultra diffuse galaxies (UDGs; [1, 2]). The discovery and characterisation of LSBGs offers an opportunity to impose constraints on theories of galaxy formation and to investigate the tensions of the concordance model of structure formation in the Universe at small scales [3]. Significant progress has been made in recent years, however, there is still much debate about the nature, formation and evolution of this type of galaxy. In particular, their total mass is a key parameter to distinguish between models of formation of these galaxies [4].

Searches of LSBGs in astronomical survey data are typically performed by selecting objects by their photometric and structural properties (e.g. surface brightness and size), which are derived from parametric modelling of objects detected in the images [5, 6, 7]. Preprocessing of the images can be employed to optimize the detection of diffuse objects or to distinguish them from background sources (e.g. [8]). Visual inspection and more refined parametric modelling are required to select

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

a pure sample of LSBGs, since a large number of low surface brightness artifacts in images, such as diffuse light from nearby bright stars or giant elliptical galaxies, often pass the selection criteria. While these techniques are needed for more precise classification of objects, they are computationally expensive and generally require large amounts of manual analysis [9].

The new generation of astronomical surveys (e.g. *Euclid*[1] and *LSST*[2]) is expected to collect hundreds of petabytes of data, and such traditional object search approaches will no longer be feasible. As the sizes of astronomical dataset grow, machine learning methods are revolutionizing the tasks of automated detection and classification of objects in images. [7], hereafter T21, used Support Vector Machines in data containing physical properties to classify LSBGs and non-LSBGs, reducing the number of objects to be modelled and visually inspected from ~ 420000 to ~ 44000 objects. However, these methods rely on a relatively large sample of both positive (LSBGs) and negative (non-LSBGs) examples, which requires substantial manual analysis due to the challenges of identifying LSBGs. This becomes a problem in new astronomical surveys that will cover large areas of the night sky. Through the usage of Self-Supervised Learning, [10] encoded images of astronomical objects as features vectors, which were then used to perform a similarity search to find gravitational lens systems. This approach was very successful, and not a single labeled example was needed during training. However, due to the usage of images, it is computationally costly and loses information related to the physical properties of objects, which are very useful when searching for LSBGs.

Thus, in this work, we propose a way of using tabular data with properties of objects detected in a large astronomical survey to automatically find new LSBG candidates with the usage of at least one known LSBG. This approach can also be used with a small sample of different LSBGs to recover a wide range of LSBG candidates.

2 Methodology

2.1 Similarity search

Similarity search can be described as a problem of quantifying how similar items stored in a dataset are to a *key* object [11]. Rigorously, one can define the problem as: let $\mathcal{D}$ be a domain, d a distance measure and $(\mathcal{D}, d)$ a metric space, given a dataset $X \subseteq \mathcal{D}$ of n elements each with m features, and an item q (key), return the k items closest to q as measured by d. This type of similarity search is called k-nearest neighbor search (kNN).

The distance function d define the way that we measure the proximity of objects in a domain. For our study, we use the Euclidean distance due to its strong applicability to data with real values [12]. This distance can be defined for m-dimensional vectors as:

$$d = [(a_1, a_2, ..., a_m), (b_1, b_2, ..., b_m)] = \sqrt{\sum_{i=1}^{m} |a_i - b_i|^2}\ , \tag{1}$$

where $(a_1, a_2, ..., a_m)$ and $(b_1, b_2, ..., b_m)$ are features. kNN can be formalized through the definition of a range function $R(q, r)$ that returns all of the objects within a distance r to the key q as $R(q, r) = \{x \in X,\ d(q, x) \leq r\}$. From that, we have that kNN is defined as:

$$kNN(q, k) = \{R \subseteq X, |R| = k \wedge \forall\ x \in R,\ y \in X - R :\ d(q, x) \leq d(q, y)\}\,. \tag{2}$$

When two or more objects share the same distance to the key, the order at which they are returned is chosen arbitrarily. Thus, kNN involves the comparison of object q with all other objects in the dataset, such that in the worst scenario computational time is given by $\mathcal{O}(nm)$ [13]. This makes the algorithm computationally expensive, specially for large dimensional data [12].

2.2 Locality-Sensitive Hashing

Fortunately, for many applications it is not essential to have the absolute nearest neighbors, and we can use approximations that make similarity search less computationally demanding. The technique

[1] `https://www.euclid-ec.org/`

[2] `https://www.lsst.org/`

we employ in our work is *Locality-Sensitive Hashing* (LSH), which uses hash functions to group data points together in buckets. Afterwards, during inference, the model only needs to perform distance calculations for the items in the same bucket as the key q [14].

The goal of LSH is to obtain a hash function such that similar items occupy the same bucket and dissimilar items occupy different buckets. Different distance functions have different hash functions that preserve locality between items. Since we are employing the Euclidean distance (Equation 1), we use t hash tables that are treated as vectors, created with random orientations, that slice the metric space. These hash tables are partitioned in equal buckets of size s. Points in the metric space are projected in each of the t tables, being hashed to the bucket that corresponds to its projection in the table. The hash functions employed are defined by [12]:

$$h_i(\vec{x}) = \left\lfloor \frac{\vec{x} \cdot \vec{v_i}}{s} \right\rfloor, \quad i = 1, 2, ..., t, \tag{3}$$

where $\vec{x}$ is a point in the $(\mathcal{D}, d)$ metric space and $\vec{v_i}$ is a hash table. Each point will be projected to every hash table according to its orientation. On inference, we iterate over each table, calculating the distance from our key to the points only for the points in the same bucket as the key. This enables us to recover the k nearest neighbors to q in a computationally efficient manner.

3 Experiments

Data For our work we use data from the Dark Energy Survey (DES)[3], more specifically, we use the *DES Y3 Gold coadd (v2.2)* catalog, which contains physical properties of objects detected in this surveyNos re [15]. Following the work of T21, we perform the same initial selection cuts to remove objects that are clearly not LSBGs from our sample. Additionally, we remove objects with `NULL` values and remove coordinate and flag features, except for the object's ID, right ascension (RA) and declination (DEC). Our selection cuts are available at `https://github.com/zysymu/lsh-astro`. Our final sample consists of 11670190 objects and 354 features, from a catalog of ~ 400 million objects. For our LSBG keys we use the catalog of T21, which, after applying the same selection criteria, contains 18685 LSBGs.

Preprocessing To process our raw data we use *PySpark*, a *Python* interface to *Apache Spark*[4], an engine that enables the programming of clusters with implicit data parallelism. In order to prepare our features, we exclude the object's ID, RA and DEC due to the fact that they are only used to locate the objects. Afterwards, we apply One-Hot Encoding [16] to the 6 categorical features in our sample. Following tabular data best practices, we normalize our data by subtracting the mean and dividing by the standard deviation of each parameter, for each item.

Model In order to perform our similarity search, we use *PySpark*'s implementation of LSH, that presents 2 hyperparameters: s and t (see Equation 3). Due to the lack of measurements of absolute distances between the objects, we cannot determine the optimal values of the hyperparameters. We tested different values of s and t and visually analyzed the nearest neighbors to different keys, setting $s = 2.0$ and $t = 3$. To perform our search, we use an *Amazon EMR*[5] cluster composed of 1 `c5.xlarge` master node and 1 `c5.4xlarge` core node. In this configuration, the LSH model is fitted in 0.7 second and the time to return the 25000 nearest neighbors to a key is ~ 6 minutes.

Results To test the performance of our model we randomly selected 10 LSBGs from T21 to be used as keys. Then, we searched for the 25000 closest objects to each key. In Figure 1 we can see one of these keys and some of the returned neighbors.

Figure 2 shows a histogram of the number of objects returned by our similarity search with respect to their distances for one key. In it we also consider which of the returned objects are present in T21. In this case, with just one LSBG (our key), searching for the 25000 nearest neighbors we manage to find 12260 galaxies from the T21 LSBG sample. Thus, using this key, at least $\sim 49\%$ of neighbors

[3] `https://www.darkenergysurvey.org/`; all of the data collected is publicly available.

[4] `https://spark.apache.org/`

[5] `https://aws.amazon.com/emr/`

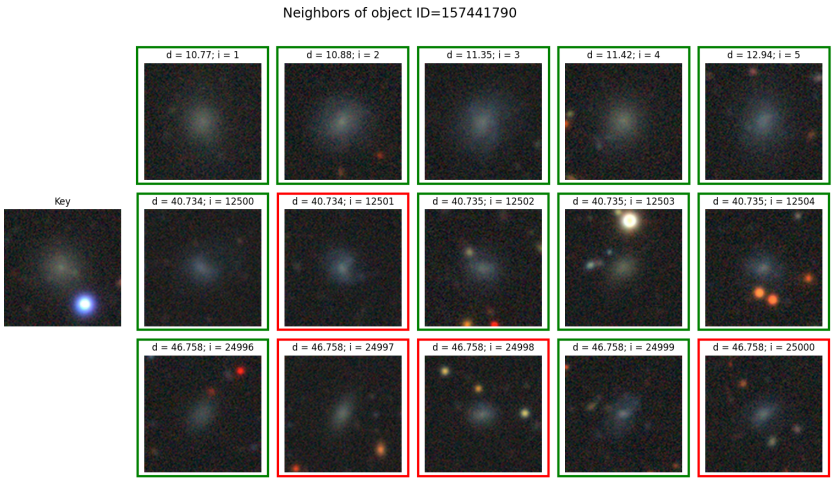

Figure 1: Closest neighbors to key object `ID=157441790`. Here, d is the distance to the key and i is the position of the neighbor. The top row has the 5 nearest neighbors, the middle row has the 5 "middle" neighbors and the bottom row has the 5 fartest neighbors. Items marked in red are not present in T21 and items marked in green are present in the catalog.

are LSBGs. Other neighbors not present in T21 are also visually very similar to the labeled LSBGs, demonstrating how our method allows for the creation of a complete sample of LSBG candidates.

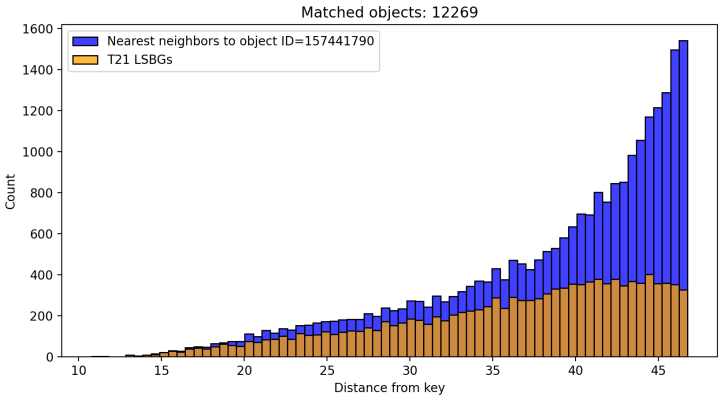

Figure 2: In blue, a distance histogram of the 25000 nearest neighbors to the LSBG `ID=157441790`. In orange, the distance histogram of the closest neighbors to the same key that are present in T21.

4 Conclusion

With future astronomical surveys expected to produce enormous amount of data, traditional ways of analyzing and generating a sample manually become unfeasible. Thus, in this work we used an approximate similarity search method, Locality-Sensitive Hashing, with tabular data that describes properties of objects from the Dark Energy Survey to create a model that allows for the identification of objects by similarity. This enables researchers to efficiently find astronomical objects of the same class with just a small labeled sample. We used this model to find several already cataloged LSBGs and objects that are visually extremely similar to other galaxies of this type but not present in catalogs. Also, due to the generality of models based on similarity searches, our method is able to return similar entries to other kinds of astronomical objects, not being limited to LSBGs.

References

[1] Pieter G. van Dokkum, Roberto Abraham, Allison Merritt, Jielai Zhang, Marla Geha, and Charlie Conroy. FORTY-SEVEN MILKY WAY-SIZED, EXTREMELY DIFFUSE GALAXIES IN THE COMA CLUSTER. *ApJL*, 798(2):L45, January 2015. Publisher: American Astronomical Society.

[2] van der Burg, Remco F. J., Muzzin, Adam, and Hoekstra, Henk. The abundance and spatial distribution of ultra-diffuse galaxies in nearby galaxy clusters. *A&A*, 590:A20, 2016.

[3] James S. Bullock and Michael Boylan-Kolchin. Small-scale challenges to the cdm paradigm. *Annual Review of Astronomy and Astrophysics*, 55(1):343–387, 2017.

[4] Michael A. Beasley, Aaron J. Romanowsky, Vincenzo Pota, Ignacio Martin Navarro, David Martinez Delgado, Fabian Neyer, and Aaron L. Deich. An Overmassive Dark Halo around an Ultra-diffuse Galaxy in the Virgo Cluster. , 819(2):L20, March 2016.

[5] D J Prole, J I Davies, O C Keenan, and L J M Davies. Automated detectionof very low surface brightness galaxiesin the Virgo cluster. *Monthly Notices of the Royal Astronomical Society*, 478(1):667–681, July 2018.

[6] D. J. Prole, R. F. J. van der Burg, M. Hilker, and J. I. Davies. Observational properties of ultra-diffuse galaxies in low-density environments: field UDGs are predominantly blue and star forming. *Monthly Notices of the Royal Astronomical Society*, 488:2143–2157, September 2019. ADS Bibcode: 2019MNRAS.488.2143P.

[7] D. Tanoglidis, A. Drlica-Wagner, K. Wei, T. S. Li, J. Sánchez, Y. Zhang, A. H. G. Peter, A. Feldmeier-Krause, J. Prat, K. Casey, A. Palmese, C. Sánchez, J. DeRose, C. Conselice, L. Gagnon, T. M. C. Abbott, M. Aguena, S. Allam, S. Avila, K. Bechtol, E. Bertin, S. Bhargava, D. Brooks, D. L. Burke, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, C. Chang, M. Costanzi, L. N. da Costa, J. De Vicente, S. Desai, H. T. Diehl, P. Doel, T. F. Eifler, S. Everett, A. E. Evrard, B. Flaugher, J. Frieman, J. García-Bellido, D. W. Gerdes, R. A. Gruendl, J. Gschwend, G. Gutierrez, W. G. Hartley, D. L. Hollowood, D. Huterer, D. J. James, E. Krause, K. Kuehn, N. Kuropatkin, M. A. G. Maia, M. March, J. L. Marshall, F. Menanteau, R. Miquel, R. L. C. Ogando, F. Paz-Chinchón, A. K. Romer, A. Roodman, E. Sanchez, V. Scarpine, S. Serrano, I. Sevilla-Noarbe, M. Smith, E. Suchyta, G. Tarle, D. Thomas, D. L. Tucker, A. R. Walker, and DES Collaboration. Shadows in the Dark: Low-surface-brightness Galaxies Discovered in the Dark Energy Survey. *ApJS*, 252(2):18, January 2021.

[8] Dennis Zaritsky, Richard Donnerstein, Arjun Dey, Jennifer Kadowaki, Huanian Zhang, Ananthan Karunakaran, David Martínez-Delgado, Mubdi Rahman, and Kristine Spekkens. Systematically Measuring Ultra-diffuse Galaxies (SMUDGes). I. Survey Description and First Results in the Coma Galaxy Cluster and Environs. *ApJS*, 240(1):1, December 2018. Publisher: American Astronomical Society.

[9] James Pearson, Jacob Maresca, Nan Li, and Simon Dye. Strong lens modelling: comparing and combining Bayesian neural networks and parametric profile fitting. *Monthly Notices of the Royal Astronomical Society*, 505(3):4362–4382, June 2021. arXiv:2103.03257 [astro-ph].

[10] George Stein, Jacqueline Blaum, Peter Harrington, Tomislav Medan, and Zarija Lukic. Mining for Strong Gravitational Lenses with Self-supervised Learning. *The Astrophysical Journal*, 932(2):107, June 2022. arXiv:2110.00023 [astro-ph].

[11] Pavel Zezula, Giuseppe Amato, Vlastislav Dohnal, and Michal Batko. *Similarity Search - The Metric Space Approach*, volume 32. January 2006.

[12] Jingdong Wang, Heng Tao Shen, Jingkuan Song, and Jianqiu Ji. Hashing for Similarity Search: A Survey, August 2014. arXiv:1408.2927 [cs].

[13] Padraig Cunningham and Sarah Jane Delany. k-Nearest Neighbour Classifiers: 2nd Edition (with Python examples). April 2020.

[14] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. *Mining of Massive Datasets*. Cambridge University Press, 3 edition, 2020.

[15] I. Sevilla-Noarbe, K. Bechtol, M. Carrasco Kind, A. Carnero Rosell, M. R. Becker, A. Drlica-Wagner, R. A. Gruendl, E. S. Rykoff, E. Sheldon, B. Yanny, A. Alarcon, S. Allam, A. Amon, A. Benoit-Lévy, G. M. Bernstein, E. Bertin, D. L. Burke, J. Carretero, A. Choi, H. T. Diehl, S. Everett, B. Flaugher, E. Gaztanaga, J. Gschwend, I. Harrison, W. G. Hartley, B. Hoyle, M. Jarvis, M. D. Johnson, R. Kessler, R. Kron, N. Kuropatkin, B. Leistedt, T. S. Li, F. Menanteau, E. Morganson, R. L. C. Ogando, A. Palmese, F. Paz-Chinchón, A. Pieres, C. Pond, M. Rodriguez-Monroy, J. Allyn Smith, K. M. Stringer, M. A. Troxel, D. L. Tucker, J. de Vicente, W. Wester, Y. Zhang, T. M. C. Abbott, M. Aguena, J. Annis, S. Avila, S. Bhargava, S. L. Bridle, D. Brooks, D. Brout, F. J. Castander, R. Cawthon, C. Chang, C. Conselice, M. Costanzi, M. Crocce, L. N. da Costa, M. E. S. Pereira, T. M. Davis, S. Desai, J. P. Dietrich, P. Doel, K. Eckert, A. E. Evrard, I. Ferrero, P. Fosalba, J. García-Bellido, D. W. Gerdes, T. Giannantonio, D. Gruen, G. Gutierrez, S. R. Hinton, D. L. Hollowood, K. Honscheid, E. M. Huff, D. Huterer, D. J. James, T. Jeltema, K. Kuehn, O. Lahav, C. Lidman, M. Lima, H. Lin, M. A. G. Maia, J. L. Marshall, P. Martini, P. Melchior, R. Miquel, J. J. Mohr, R. Morgan, E. Neilsen, A. A. Plazas, A. K. Romer, A. Roodman, E. Sanchez, V. Scarpine, M. Schubnell, S. Serrano, M. Smith, E. Suchyta, G. Tarle, D. Thomas, C. To, T. N. Varga, R. H. Wechsler, J. Weller, and R. D. Wilkinson. Dark Energy Survey Year 3 Results: Photometric Data Set for Cosmology. *ApJS*, 254(2):24, May 2021. Publisher: American Astronomical Society.

[16] Sebastian Raschka and Vahid Mirjalili. *Python Machine Learning, 3rd Ed.* Packt Publishing, Birmingham, UK, 2019.