Similarity Search of Low Surface Brightness Galaxies in Large Astronomical Catalogs

Marcos Tidball*, Cristina Furlanetto

Instituto de Fisica, Federal University of Rio Grande do Sul

[*marcostidball@gmail.com]

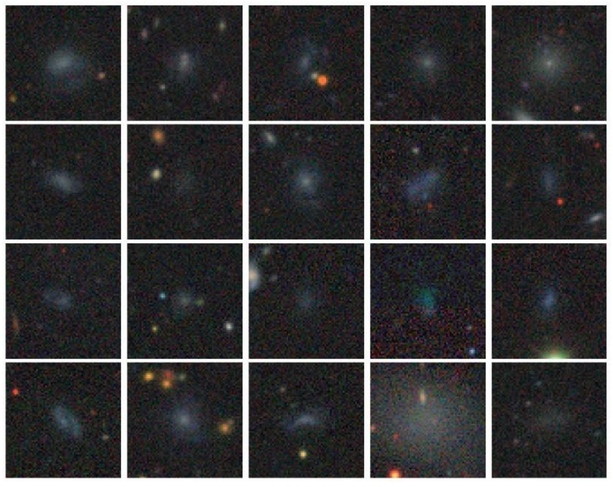
Search for Low Surface Brightness Galaxies (LSBGS)

LSBGs constitute an important segment of the Universe's galaxy population, with a significant part of the matter in the local Universe being thought to be "hidden" in such diffuse objects. However, due to their low surface brightness, their identification is challenging [1]. Searches for LSBGs are usually performed with a combination of parametric modelling and visual inspection, which are required for a pure sample of LSBGs but become unfeasible on large scales.

Machine learning methods based on tabular data have been successful in reducing the number of objects to be modelled and visually inspected, with the work of [2], hereafter T21, reducing a sample of ~420000 objects to ~44000 LSBG candidates. However, traditional supervised methods require a large sample of positive and negative examples, which becomes a problem due to the difficulty of detecting LSBGs.

Our contribution

In this work we propose the usage of an approximate nearest neighbors algorithm with tabular data of object properties to automatically find new LSBG candidates needing just one known LSBG. We intend this approach to be used with a small sample of different LSBGs to recover a wide range of LSBG candidates with at least one labeled example.



Examples of LSBGs from T21 [2].

Locality-Sensitive Hashing (LSH)

To perform approximate similarity search we employ LSH, which uses hash functions such that similar items occupy the same bucket and dissimilar items occupy different buckets. To measure similarity, we use the Euclidean distance, which uses the following family of hash functions [3]:

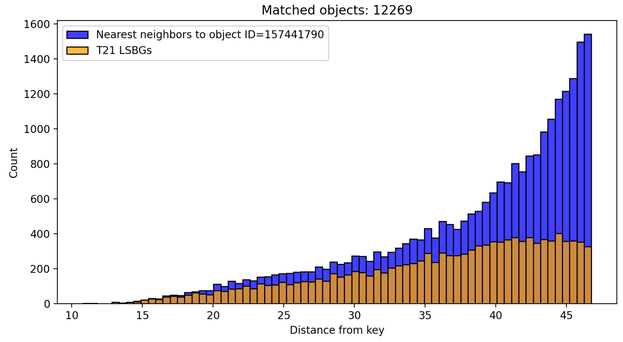
$$h_i(\vec{x}) = \left\lfloor \frac{\vec{x} \cdot \vec{v_i}}{s} \right\rfloor, \quad i = 1, 2, ..., t,$$

t is the number of hash tables, which are partitioned in equal buckets of size s. Points in the metric space ($\vec{x}$), are projected to each table ($\vec{v_i}$), being hashed to a bucket. Each point is projected according to the orientation of each hash table.

We perform a k-nearest neighbor search. During inference LSH iterates over each hash table, calculating the distance from the key object to the points hashed in the same bucket as the key. This enables us to recover the nearest neighbors in an efficient manner for large dimensional data.

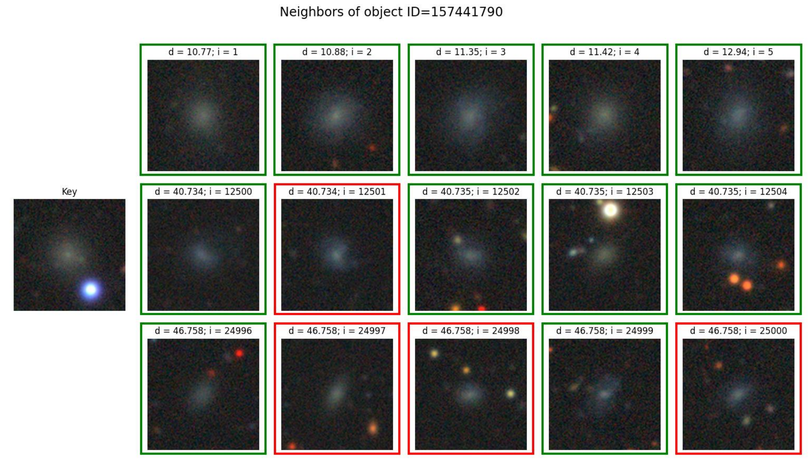
Data and experimental setup

We use the *DES Y3 Gold coadd (v2.2)* catalog [4], which contains physical properties of astronomical objects. We perform a combination of selection cuts done in T21 (to remove objects that are clearly not LSBGs) and quality cuts. Our final sample consists of 11670190 objects and 354 features. For our LSBG keys we use the catalog of T21, which, after the same cuts, contains 18685 LSBGs. We use *PySpark* to process our data and train our model. We One-Hot Encode categorical features and normalize our data. We set s = 2.0 and t = 3.0.



Results and conclusion

To test the performance of our model we randomly selected 10 LSBGs from T21 to be used as keys and searched for the 25000 nearest neighbors of each key. We manage to recover a large portion of the T21 catalog while needing only 1 labeled LSBG example. Also, other neighbors that are not present in T21 are visually very similar to the key, demonstrating how our method is capable of finding LSBG candidates not present in current catalogs.



Objects marked in green are LSBGs present in T21; objects marked in red are not present in T21.

Future astronomical surveys are expected to produce enormous amounts of data, making traditional ways of manually analyzing and generating a sample unfeasible. In this work we used LSH with tabular data to create a model that allows for the identification of astronomical 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 catalogued LSBGs and objects that are visually very similar to LSBGs but are not present in the T21 catalog.

References

[1] P van Dokkum et al. (2014). "Forty-Seven Milky Way-Sized, Extremely Diffuse Galaxies in the Coma Cluster". In: ApJL, 798:L45.

[2] D. Tanoglidis, et al. (2021). "Shadows in the Dark: Low-surface-brightness Galaxies Discovered in the Dark Energy Survey". In: ApJS, 252:18.

[3] J. Wang et al. (2014). " Hashing for Similarity Search: A Survey". In: arXiv:1408.2927. Bart Simpson (2013). "Hello". In: IEEE Simpsons.

[4] I. Sevilla-Noarbe et al. (2021). "Dark Energy Survey Year 3 Results: Photometric Data Set for Cosmology". In: ApJS, 254:24

GitHub