Nearest Neighbour Regression Outperforms Model-based Prediction of Specific Star Formation Rate

Kristoffer Stensbo-Smidt*, Christian Igel*, Andrew Zirm†, Kim Steenstrup Pedersen*
*Department of Computer Science, University of Copenhagen
†Dark Cosmology Centre, Niels Bohr Institute, University of Copenhagen

Abstract—Data in astronomy is rapidly growing with upcoming surveys producing 30 TB of images per night. Highly informative spectra are too expensive to measure for each detected object, hence ways of reliably estimating physical properties from images alone are paramount. The objective of this work is to test whether a “big data ready” k-nearest neighbour regression can successfully estimate the specific star formation rate (sSFR) from colours of low-redshift galaxies. The nearest neighbour algorithm achieves a root mean square error (RMSE) of 0.30, outperforming the state-of-the-art astronomical model achieving a RMSE of 0.36.

Keywords—machine learning, astronomy, nearest neighbour regression, specific star formation rate

I. Introduction

The traditional workflow in astronomy has heavily relied on the analysis of single events. Acquiring and processing data was slow and expensive and the resulting post-processed data were often of low quality. In the past decades, however, ever larger astronomical surveys, such as the Sloan Digital Sky Survey (SDSS), have appeared, which have led to an enormous amount of high-resolution and publicly available astronomical images. The SDSS alone provides images of nearly $10^9$ objects, making astronomy enter a new big data era. Even larger surveys will start collecting data soon. The Large Synoptic Survey Telescope (LSST) is expected to produce 30 TB of images per night, which need to be analysed in near real-time, in order to trigger time-critical follow-up observations. Thus, efficient large-scale data mining techniques are required.

Many astronomically interesting properties of an object can only be measured from a spectrum of its light. A spectrum is, roughly speaking, the photon count at each wavelength of the received light, and since each chemical element has a unique spectrum, a lot of information can be extracted from the light of, for instance, a star or a galaxy (see figure 1). Spectra of stars are very well understood, but for galaxies containing stars, gas and dust, each contributing to the final spectrum, extracting relevant information is more difficult. To extract such information, the individual contributions from stars, gas and dust must be disentangled. To this end, astronomers build models. These models simulate emission of light from stars from different populations at different ages, adding in the contributions from gas and dust in the galaxy and in the intermediate space. Comparing the resulting model spectra to the real spectrum then allows astronomers to extract relevant information about the galaxy.

One particular galactic property that we are concerned with in this study is the rate at which new stars are formed, the star formation rate (SFR). The SFR is an important quantity to measure, as it relates to the evolutionary history of a galaxy, and it is most easily measured from the spectrum of the galaxy. A derived measure often used by astronomers is the specific star formation rate (sSFR), which is the SFR normalised by the mass of stars in the galaxy.

Observation time is, unfortunately, a scarce resource, and measuring a spectrum is very expensive and time consuming. A much cheaper way to study the sky is to take images. These images can contain thousands of objects and are usually acquired in multiple bandpass filters, each spanning a range of the spectrum, see figure 2. From these images, it is possible to extract the intensity of an object in each filter, and from the intensities, one can extract the magnitude of the object in each filter. Subtracting two magnitudes from each other effectively removes a bias parameter, and the resulting value is known as a colour. The colours of an object can also be found from the spectrum itself; the intensity of the object in a given filter is the product of the filter transmission with the spectrum, integrated over all wavelengths. The colours of an object are therefore linked to its spectrum and so should contain a significant amount of information.

Against this background, we study the potential of using a scalable machine learning approach to predict the sSFR from colour values obtained from galaxy images. The next section briefly describes the data and the current model-based approach to sSFR calculation and prediction. We then describe our machine learning approach. Sections IV and V present experiments and results before we conclude.

Figure 1. Examples of well-resolved galaxies in the SDSS database.
The data considered in this study come from Sloan Digital Sky Survey (SDSS), an astronomical survey that has been scanning the night sky since 2000, see figure 1. The SDSS provides photometric data (images) as well as spectroscopic data (spectra) of selected objects. The photometric data are observed in five bandpass filters, spanning wavelengths from 3000 Å to 11 000 Å, while the spectroscopic data span wavelengths from 3800 Å to 9200 Å, see figure 2. The database currently contains images of $9.3 \times 10^8$ objects and spectra of $1.6 \times 10^6$ objects, about half being galaxies.

Ground truth sSFR values for selected objects are also made available by SDSS. The derivation of these values is described in the following, see [1] and [2] for details. The SFR of a galaxy is measured from parts of its spectrum associated with newly formed stars. These parts, however, are contaminated by contributions from both inter- and intragalactic gas and dust, and a disentangling of these components is necessary. To do so, a library of $\sim 2 \times 10^5$ artificial model spectra are generated from simulations with varying physical parameters. From these spectra the SFR and stellar mass, and in turn the sSFR, can be calculated. To estimate the sSFR of a real galaxy, its spectrum is compared to each of the model spectra. The sSFRs of the models are then weighted based on the likelihood of the model magnitudes given the real spectrum, resulting in a probability distribution for the sSFR. From this distribution, the final sSFR of the galaxy is calculated as the expected sSFR.

If only photometric data are available, estimation of the sSFR must rely on the magnitude of the galaxy in each bandpass filter. The current state-of-the-art approach in this regard (see [3]) follows the procedure outlined above to a large extent; a library of model spectra, each associated with a particular set of physical parameters, is generated, but instead of using the spectra directly, they are multiplied with the SDSS filter transmissions and integrated over all wavelengths to produce model magnitudes. To estimate the sSFR of a real galaxy, its photometric magnitudes are compared to the magnitudes of each model in the library. The sSFRs of the models are then weighted based on the likelihood of the model magnitudes given the real magnitudes, resulting in a probability distribution for the sSFR. From this distribution, the predicted sSFR of the galaxy is again calculated as the expected sSFR.

Data used in this study: The data used in this study are a subset of SDSS Data Release 7 (DR7). For each galaxy, only the five magnitudes (we use the SDSS MODEL MAG magnitudes) and the sSFR are considered. The magnitudes have been subtracted from each other, resulting in a four-vector representing the colours. We only consider galaxies in SDSS that meet the following selection criteria:

- A spectrum is available, that is, a reliable ground truth sSFR is available.
- Magnitudes are larger than $-9999$ (which indicates a failed magnitude estimation).
- sSFR is larger than $-99$ (which indicates a failed sSFR estimation).

A sample of 694 894 galaxies meet these criteria. For a small subset of 11 450 low-redshift galaxies within the selected sample, we additionally have sSFR predictions obtained by the template-based modelling approach described above. No additional selection criteria have been applied to this subset.

III. PREDICTING THE SSFR USING SIMPLE MACHINE LEARNING TECHNIQUES

Measuring a spectrum of a galaxy is a time-consuming task, so in order to take advantage of the enormous amounts of readily available astronomical data, it is essential that one can reliably estimate physical properties from photometric data alone. The template-based method, as described in section II, is, however, not a viable solution considering the rapidly growing amount of data. Instead, scalable machine learning techniques provide a promising solution to this problem.

The large data sets in astronomy require fast algorithms, and it is therefore common to resort to very basic machine learning techniques. For example, nearest neighbour regression has been successfully applied to mining SDSS data for detecting distant quasars, one of the most challenging tasks in astronomy [4].

In this work we use exact $k$ nearest neighbour regression ($k$-NN, see [5] for an overview of efficient, scalable implementations) to estimate the sSFRs for a set of galaxies, using only their four colours computed from five photometric magnitudes. For a given input sample $x \in \mathbb{R}^4$, the $k$ nearest neighbours within the training data are determined using the Euclidean distance. The average sSFRs of the $k$ neighbours is the predicted sSFR of $x$. The dimensionality of the feature space...
space is very low and the training data set is sufficiently large making $k$-NN regression a viable approach.

IV. EXPERIMENTS

We performed two experiments, one comparing the machine learning with astrophysical modelling results and another assessing the approach on a larger subset of the SDSS database. These experiments were conducted using the Shark machine learning library [6].

A. Comparison with template-based model

We compared the nearest neighbour approach with the template-based model described above on the subset of 11 450 galaxies for which results from the template-based model are available. To assess the performance of the algorithm, we used cross validation (CV) partitioning the data into 10 folds [7]. We refer to this procedure as the outer CV.

For each of the 10 folds, we independently determined a proper value of $k$. This was done using an inner 10-fold CV procedure splitting the available training fold from the outer CV. We selected the $k \in \{1, 3, 5, \ldots, 25\}$ with the lowest average inner CV test error. This $k$ was used to predict the sSFRs of the outer CV’s validation data fold using the outer CV’s training data fold.

To make the estimations by the $k$-NN and the template-based model comparable, the predictions by the template-based model were divided into the same 10 subsets as used in the outer CV of the $k$-NN and the same statistics were calculated.

B. Application to larger subset of SDSS

In the previous experiments, the $k$ values chosen in the 10 model selection procedures were concentrated around $k = 21$, however, different values of $k$ gave similar performance as long as $k > 15$. In the second experiment, we considered our full SDSS DR7 subset, for which spectrum-based ground truth sSFRs are available. We randomly selected 30 000 galaxies for training a 21-NN regression and used the remaining 664 894 galaxies for testing.

V. RESULTS

For the first experiment, we report the average RMSEs and the standard deviations\(^1\) over the 10 validation sets of the outer CV of the predictions by the $k$-NN regression and the state-of-the-art template-based model compared to the ground truth sSFRs. The template-based model achieved a RMSE of $0.36 \pm 0.01$ on our data set, while the $k$-NN model achieved a RMSE of $0.30 \pm 0.01$. The $k$-NN model performs considerably better than the slower, template-based, physically motivated, model.

\(^1\)The standard deviations should not be interpreted as strict confidence intervals, since the data folds in the CV procedure are not statistically independent.

A visualisation of the correlations between the predictions of the two models and the ground truths is given in figure 3. The main difference between the predictions by the two models seems to be the position of the peak corresponding to low sSFRs; while the $k$-NN model overshoots a bit, the template-based model undershoots considerably.

In the second experiment, the $k$-NN regression also achieved a RMSE of 0.30 on the larger, more heterogeneous data set. A visualisation of the correlations between the predictions of the $k$-NN regression and the ground truths is given in figure 4.

VI. CONCLUSIONS

Increasingly larger astronomical surveys are leading to an enormous amount of high-resolution and publicly available astronomical images. Many astronomically interesting properties of an object can, however, most reliably be measured from a spectrum of its light, which is expensive to measure. A much cheaper way to make astronomical discoveries is to use images directly. Many physical properties can be
estimated from these images, though much less accurately than from a spectrum. The large data sets demand more efficient ways of estimating quantities than the current state-of-the-art model provides, if we are to take advantage of the wealth of information hidden in the data. This invites for applying scalable machine learning algorithms. We exemplified this by predicting the specific star formation rate (sSFR) of galaxies using a scalable machine learning approach, which outperformed the state-of-the-art astronomical model. On a small sample of low-redshift galaxies from the SDSS DR7 database, nearest neighbour regression achieved a RMSE of 0.30 compared to 0.36 achieved by a template-based, physically motivated model. Applied to all galaxies with spectra and reliable colour and sSFR measurements in DR7, the error of our approach was 0.30. The proposed algorithm allows for screening large amounts of imaging data and for picking out samples for further analysis. An advantage of the template-based model is, of course, the additional physical knowledge gained from simulating galaxies. The higher accuracy achieved by the $k$-NN model indicates, however, that the template-based model does not capture all the information available in the colours.

**Acknowledgements**

KSS, CI, and KSP gratefully acknowledge support from The Danish Council for Independent Research | Natural Sciences through the project “Surveying the sky using machine learning”.

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is http://www.sdss.org/

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

**References**


