Heidelberg Institute for Theoretical Studies
HITS
University of Copenhagen
Department of Computer Science
University of Copenhagen
Niels Bohr Institute, DARK

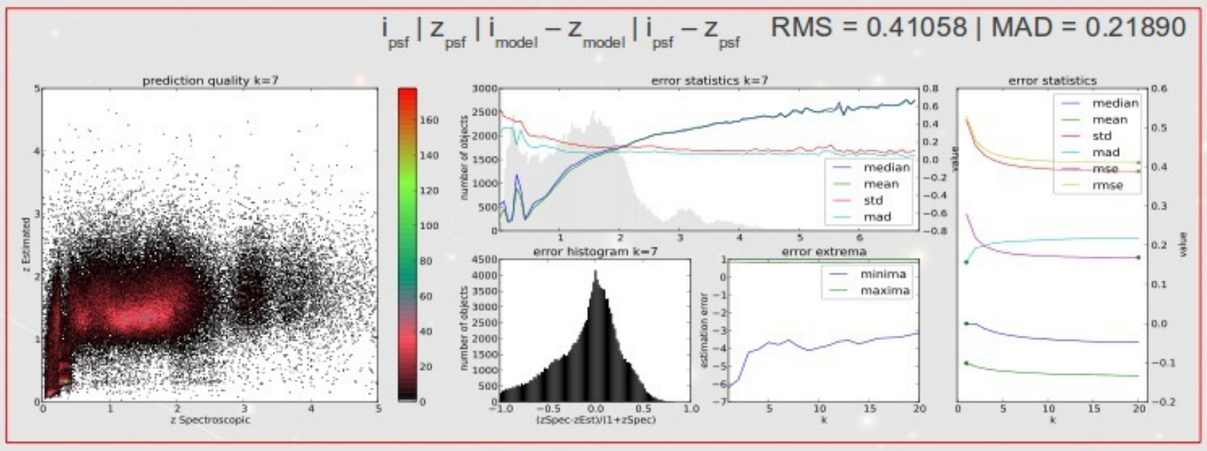
Improving the Performance of Photometric Regression Models via Massive Parallel Feature Selection

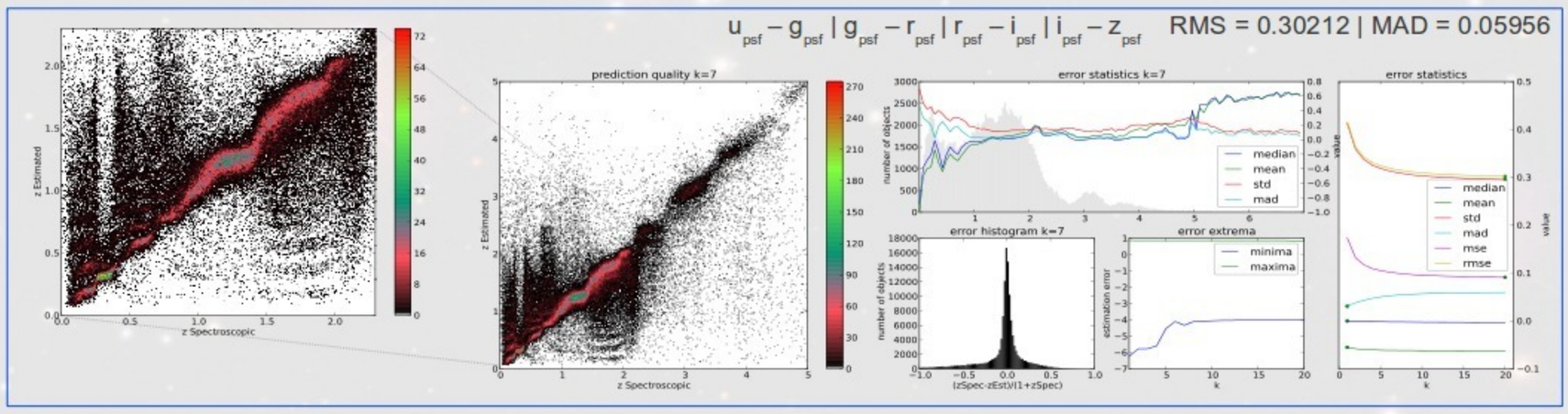
Kai Lars Polsterer[1] Kai.Polsterer@h-its.org
Fabian Gieseke[2] Fabian.Gieseke@diku.dk
Christian Igel[2] Igel@diku.dk
Tomotsugu Goto[3] tomo@dark-cosmology.dk

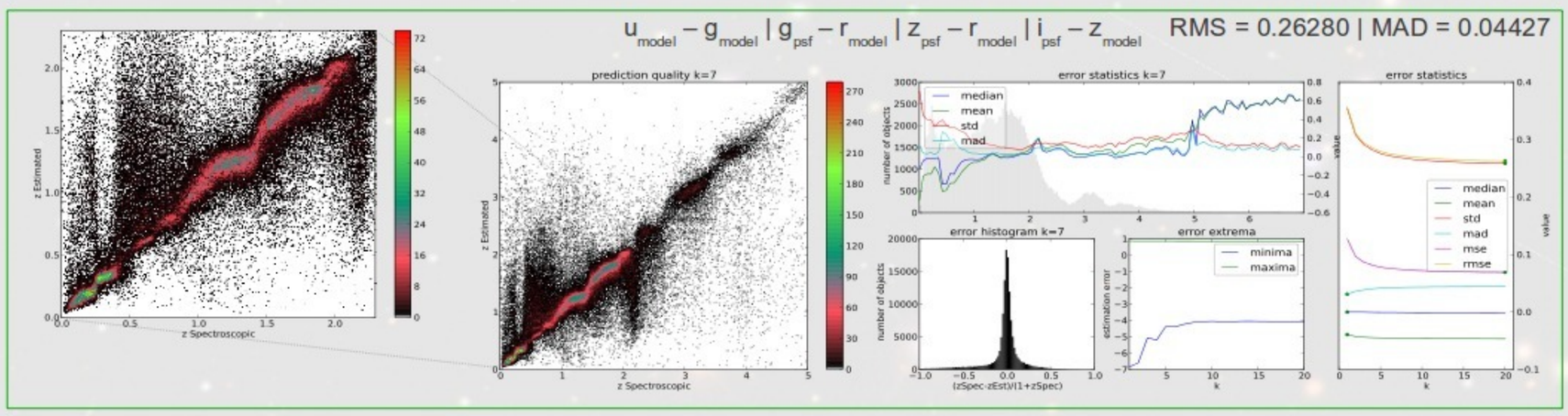
[1]Heidelberg Institute for Theoretical Studies, Astroinformatics, Heidelberg, Germany [2]University of Copenhagen, Department of Computer Science, Copenhagen, Denmark [3]University of Copenhagen, Dark Cosmology Centre, Niels Bohr Institute, Copenhagen, Denmark

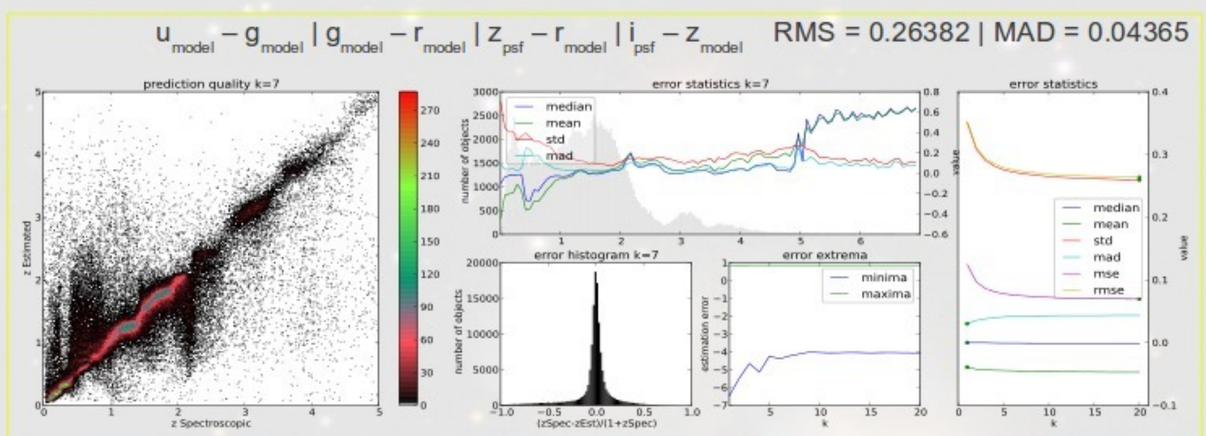
Abstract:

Regression tasks are common in astronomy. Typical problems are, for instance, the estimation of the redshift z or the metallicity of galaxies. Generating such regression models, however, is often complicated by the heterogeneity of the available input catalogs, which leads to missing data and/or features differing in explanatory power. In this work, we show how simple but effective feature selection schemes from data mining can be used to significantly improve the performance of regression models for photometric redshift and metallicity estimation (even without any particular knowledge of the physical properties of the input parameters). Our framework tests huge amounts of possible feature combinations. Since corresponding (single-core) implementations are computationally very demanding, we make use of the massive computational resources provided by nowadays graphics processing units to significantly reduce the overall runtime. This makes the method applicable to large-scale tasks in astronomy, as we demonstrate in our experimental evaluation. We conclude the work by discussing further applications of the proposed approach in astronomy.









Motivation:

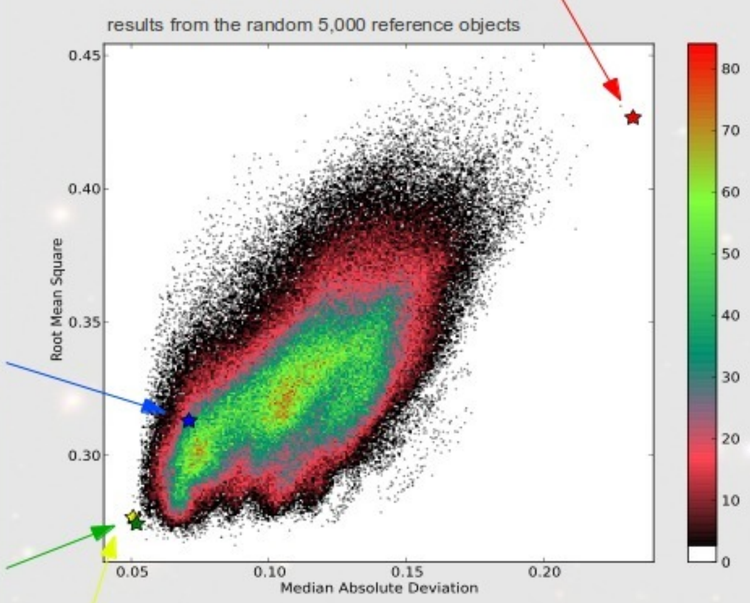
Photometric redshift estimation models depict one of the key tools in the field of astronomy. In the literature, such models are often based on the colors derived from adjacent filter bands. In this work, we consider an alternative approach: Instead of resorting to the standard features, we consider a large set of alternative though physically motivated combinations of features and present a simple yet effective feature selection scheme to determine the best performing subset of final input features (with cardinality four). As shown in the experimental part of this work, these features lead to a significantly better performance compared to the standard features usually used in astronomy.

Selecting the best-performing subset of features, however, is a combinatorial task, which quickly becomes infeasible. Our approach makes use of the massive computational resources provided by nowadays graphics processing units (GPUs). The overall approach combines this feature selection scheme with nearest neighbor models. Since the target feature space is low-dimensional, this technique depicts a very reasonable choice.

Experiment:

The photometric features of all 134,545 quasars from the SDSS with spectroscopically determined redshifts have been retrieved and locally stored for further processing. Both the point spread function (PSF) magnitudes and the model magnitudes in all 5 filter bands (u, g, r, i ,z) are used. The measurement errors and quality flags which are stored in the catalog, are ignored to create an easily reproducible sample. To be able to handle the dataset in the memory of the GPU more efficiently, a subset of 5,000 objects was selected randomly as our training and validation set. A comparison between our random sample and a sample which shows a more homogeneous density distribution in the redshift space, shows a better transferability of the performance to the complete dataset with the random sample.

Besides the plain psf and model features, we created all pair differences between the features and ended up with a total of 55 composed features. To minimise the effects of different value ranges of the individual features, which limits the prediction quality that is based on euclidean distance, we normalized all of them. A min-max, a standard deviation dependent, and a median absolute deviation (MAD) based normalisation have be evaluated, but no major differences in the prediction quality could be observed. Therefore a simple min-max normalisation was used.



Results:

To select the best r=4 features from the available n=55 features, we tested all possible n!/((n-r)!r!)=341,055 feature combinations. By plotting the RMS against the MAD we were able to pick one of the pareto optimal solutions. This feature combination shows both, a low MAD and a low RMS. What is interesting about the found feature combination is their similarity to the standard features. In most cases adjacent filter bands are used to build a color feature instead of using the raw band information. This is most probably related to the reduction of object intrinsic luminosity properties that are not related to the objects redshift. Another obvious selection effect is that mixed features are combined of psf and model magnitudes. This allows the regression model to detect the slightly extended host galaxy of nearby quasars and therefore improve the prediction quality in the low redshift regime.

Conclusion:

By using a massive parallel feature selection approach on a GPU we were able to determine the best four features to photometrically determine the redshift of quasars with a kNN regression model. This general method shows that optimizing well known models by brute force testing all possible feature combinations can lead to improved results. It even might be useful to generate feature combinations that are optimized just for one certain task, by specifying different quality measures.

Acknowledgements:

SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration. Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is http://www.sdss3.org/.