Detecting Quasars in Large-Scale Astronomical Surveys

Kai Lars Polsterer[1] Fabian Gieseke[2] Andreas Thom[2] Peter Zinn[1]
Dominik Bomans[1] Ralf-Jürgen Dettmar[1] Oliver Kramer[2] Jan Vahrenhold[2]

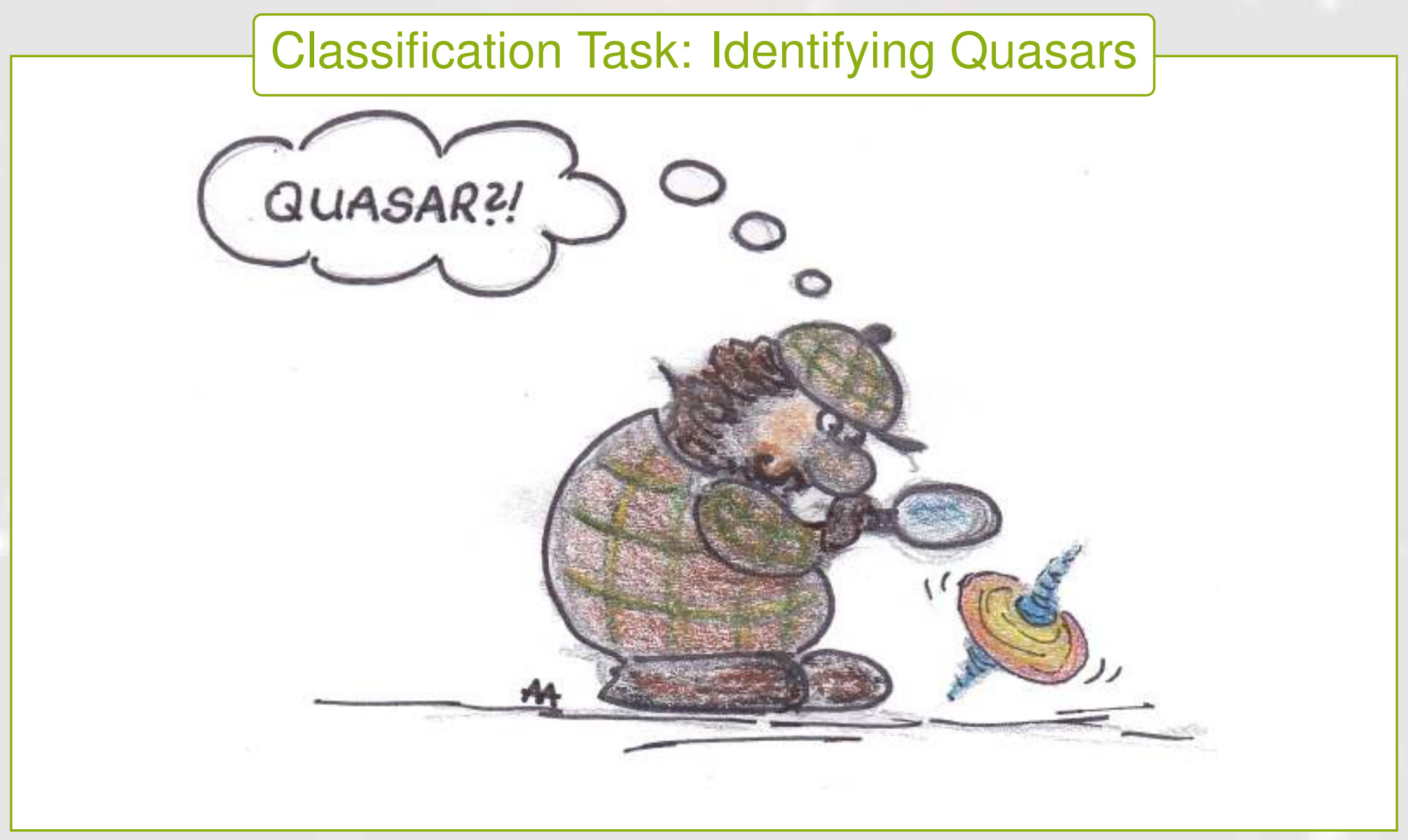
[1]Astronomisches Institut, Department of Physics and Astronomy, Ruhr-University of Bochum, Germany [2]Chair of Algorithm Engineering, Faculty of Computer Science, Technische Universität Dortmund, Germany

Abstract

We consider the task of detecting quasars in the Sloan Digital Sky Survey based on both spectroscopic and photometric data. The performances of our spectroscopic classification approaches are evaluated on a manually labeled training set. The experiments indicate that the photometric features are sufficient in order to obtain a reasonable classification performance for this particular data set and that the performance can further be improved by incorporating meaningful features obtained from the spectroscopic data.

Motivation

Classification Task: Identifying Quasars



The (semi-)automatic analysis of data sets has become an increasingly important issue for researchers in astronomy [1, 2]. This is especially true for massive data sets obtained from, e.g., the Sloan Digital Sky Survey [3] which is based on raw data of about 60 TB. From a machine learning perspective, a variety of problems in astronomy can be formulated as supervised (e.g. classification, regression) or unsupervised tasks (e.g. clustering, dimensionality reduction). We describe the use of supervised learning techniques to identify quasi-stellar radio sources (quasars) based on both spectroscopic and photometric data.

Data

Our data set is obtained from the spectroscopic data available in the Sloan Digital Sky Survey (DR6) which is said to be "one of the most ambitious and influential surveys in the history of astronomy" [3]. The data for this survey has been obtained via a 2.5-meter telescope at the Apache Point Observatory which is equipped with two special-purpose instruments: a 120-megapixel camera and a pair of spectrographs.

Apache Point Observatory (New Mexico)

Source: `http://www.sdss.org`

Classification Approach

The k-Nearest Neighbor classifier uses the k "closest" objects from the given set of classified objects to assign a class to an unclassified object [4]. More precisely, the (binary) classification $\hat{Y}(\mathbf{x})$ for an object $\mathbf{x}$ is

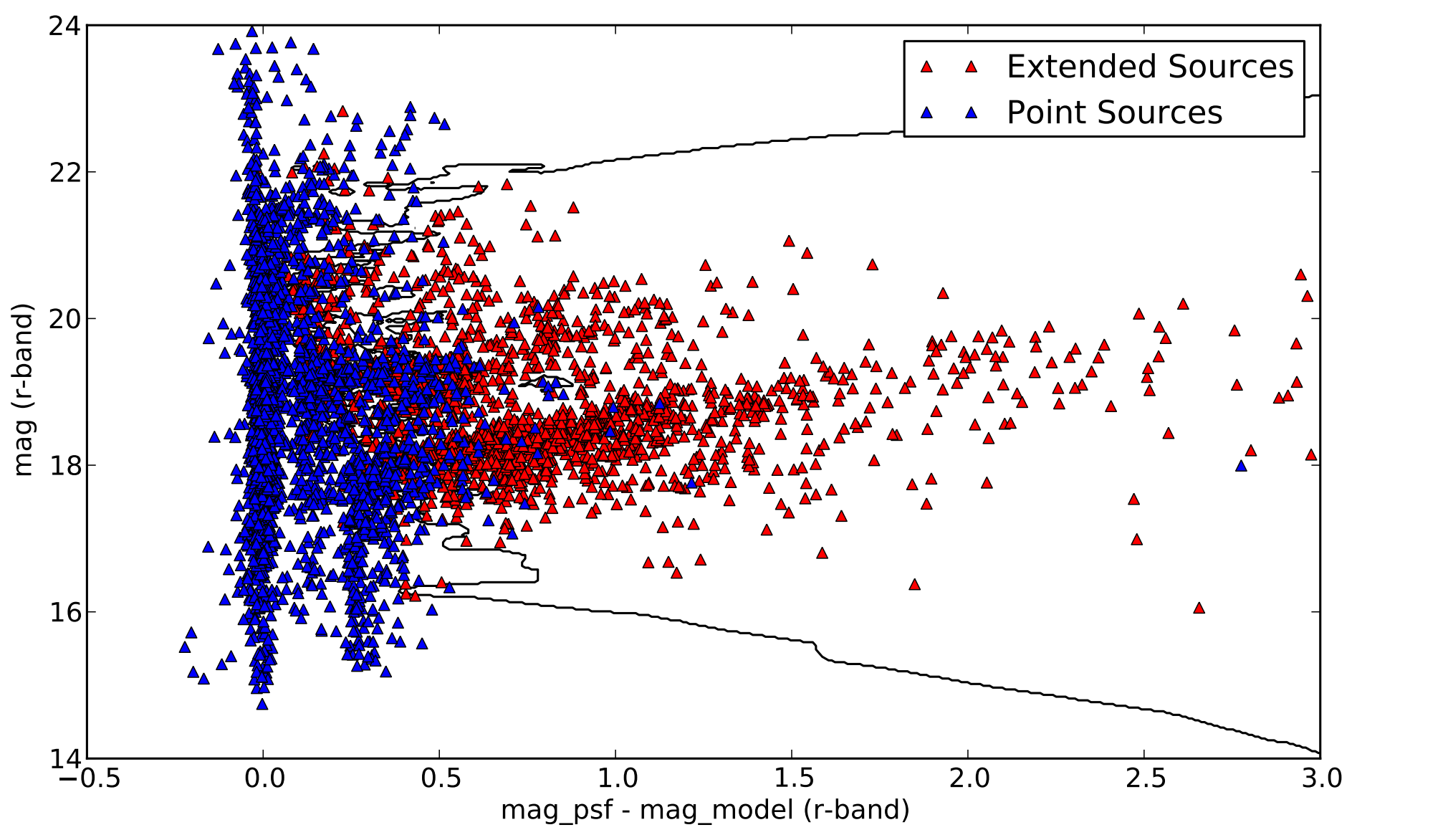
$$\hat{Y}(\mathbf{x}) = \begin{cases} 1 & \text{if } f(\mathbf{x}) > 0 \\ -1 & \text{if } f(\mathbf{x}) \leq 0, \end{cases} \tag{1}$$

where

$$f(\mathbf{x}) = \frac{1}{k} \sum_{\mathbf{x}_i \in N_k(\mathbf{x})} y_i \tag{2}$$

and where $N_k(\mathbf{x})$ denotes the k-nearest neighbors in the training set with respect to $\mathbf{x}$. To define "closeness", arbitrary metrics can be used; a popular choice is the Euclidean metric.

k-Nearest Neighbor Classifier (Selected Data Set)

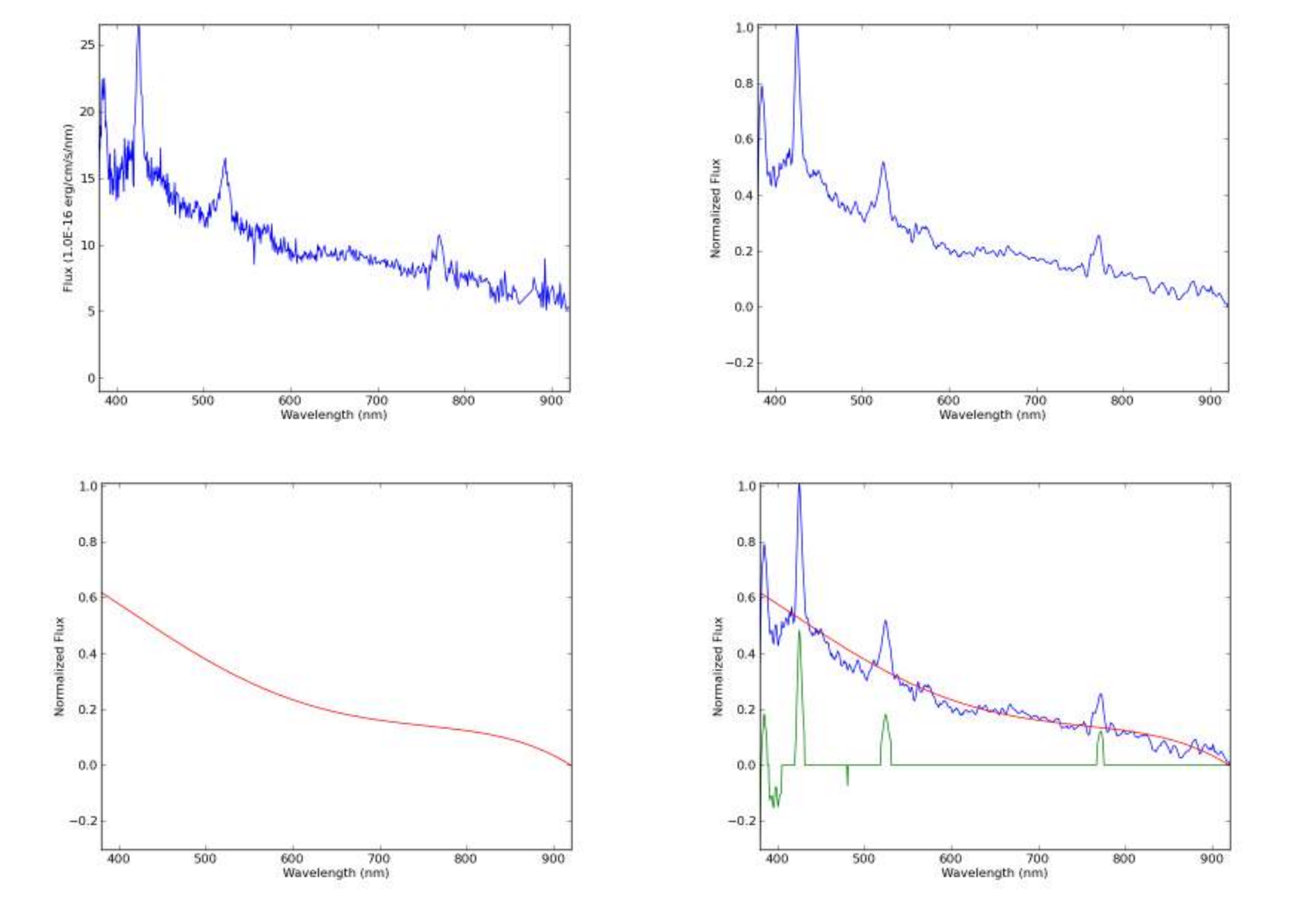


Spectroscopic Classification

Data: Given the photometric and spectroscopic data we generate a variety of data sets, see Table 2. Each data set contains all $N = 5,261$ objects ($p = 510$ objects of type "quasar" and $n = 4,751$ objects of type "other"). The labels have been obtained manually by an expert based on the spectroscopic data given for each object.
Quasars are distant galaxies with an active galactic nucleus and thus their spectroscopic data exhibits broad emission lines. Before applying our classification approaches, we thus attempt to first extract these meaningful features using the following preprocessing step:

1. Merge consecutive flux values to obtain a "binned" version.
2. Apply a smoothing filter (e.g., the Savitzky-Golay-Filter).
3. Perform a spline interpolation to extract the continuum.
4. Extract peaks based on smoothed spectrum and spline model.

Intermediate Results



The final features can then be created based on the continuum and the extracted peaks:

Feature	Description
F1	`First value of the extracted continuum`
F2	`Last value of the extracted continuum`
F3	`Integral of all positive peaks`
F4	`Integral of all negative peaks`
F5	`Width of the broadest positive peak`
F6	`Width of the broadest negative peak`
F7	`Major peak intensity`
F8	`Minor peak intensity`
F9	`Major face of positive peak`
F10	`Major face of negative peak`

Experimental Setup: To train and evaluate the k-Nearest Neighbor classification approach, half of each data set is used as training and the other half as test set. The model parameter k is determined via 10-fold cross-validation [4] on the training set.

Extracted Features: We evaluate the k-Nearest Neighbor classifier for the following features:

Data Set	Features

D1	`psfMag_u - psfMag_g, psfMag_g - psfMag_r,`
	`psfMag_r - psfMag_i, psfMag_i - psfMag_z`

D2	`BinnedSpec500`
D3	`ExtractedFeatures`

Results: Since our data sets are imbalanced, we resort to the so-called Matthews Correlation Coefficient (MCC) [5] as quality measure which takes into account the different types of errors (true positives, false positives, true negatives, and false negatives). We also provide the true positive rate (TP-rate) given by TP/p as well as the false positive rate (FP-rate) given by FP/n which can be seen as "hit rate" and "false alarm rate", respectively [6]. The classification performances of the kNN-classifier on the various data sets are shown in the following table.

Classification Performances

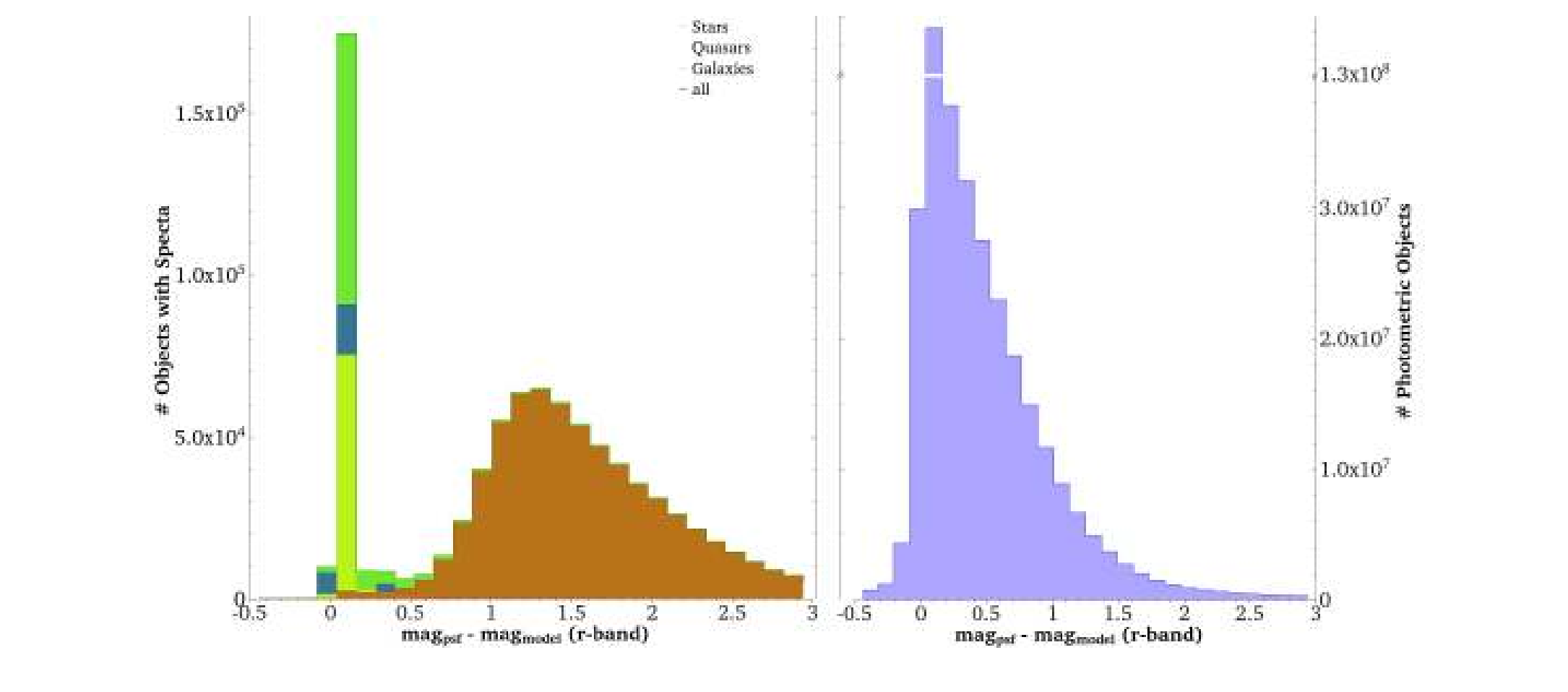
Data	kNN			
Set	MCC	Error	TP	FP
D1	0.881	2.17%	86.8%	0.9%
D2	0.810	3.38%	77.4%	1.2%
D3	0.969	0.57%	94.9%	0.0%

To test the efficiency of our approach we resort to the recently published DR7 quasar catalog [7] (considering all $N = 81,015$ quasars also given in DR6 catalog). Here we obtain a hit rate of 91.9%.

Photometric Classification

Data: The data set used for the photometric classification experiments is created by combining the photometric PSF- and model-magnitudes with the spectroscopically obtained labels. An analysis of the distribution of the r-band magnitudes indicates that the sample selection criteria of the spectroscopic targets create two major populations. In contrast to the distribution of all photometric data sets, objects with a low spatial extension are not covered accordingly. Therefore we selected a subset of the spectroscopic data that contains 82,344 objects and that reproduces the distribution of all photometric objects.

$mag_{psf} - mag_{model}$ Histogram on Spectral/Photometric Sample



Experimental Setup: In our experiments we use k-Nearest Neighbor classifiers that are trained with different features and training sets with varying sizes. The parameter k was set to 5 for all experiments.
An 8-fold cross-validation [4] is used to analyze the performance of the classifiers. On these partitions, a classifier is built with the first n elements of the training set and the classification error statistics are collected on the corresponding test set. In all experiments, false positive and false negative classifications are comparable and therefore are omitted from the plots.

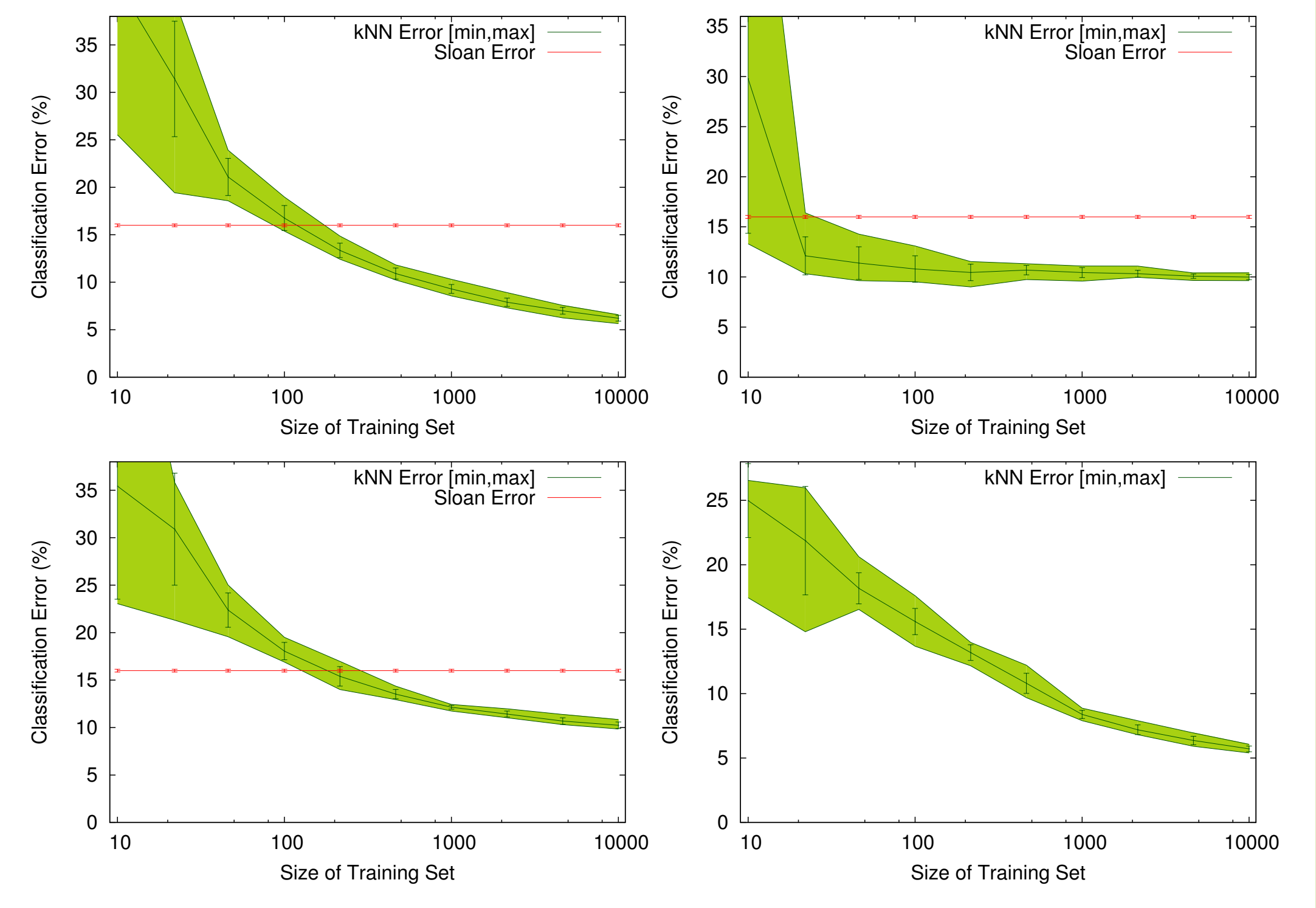
Experiment 1-3: The photometric pipeline of the Sloan Digital Sky Survey classifies objects as "extended" if the difference between the PSF- and model-magnitude in two of the three bands (g,r,i) exceeds 0.145. The SDSS classifier has an error of $\approx 16\%$ on the selected data set. Equivalent results are achieved by the k-Nearest Neighbor classifier for a training set with ≈ 200 elements. Given more elements the classifier significantly outperforms the SDSS classifier.
In the first three experiments we evaluate the quality of the point-extended-source separation with different sets of features.

- **Features 1:** mag$_{psf}$(u,r,g,i,z), mag$_{model}$(u,g,r,i,z)
- **Features 2:** mag$_{psf}$-mag$_{model}$(u,r,g,i,z)
- **Features 3:** mag$_{psf}$(u-g), mag$_{psf}$(g-r), mag$_{psf}$(r-i), mag$_{psf}$(i-z)
- **Labels 1-3: A:** stars and quasars, **B:** galaxies

Experiment 4: When trained with 10,000 data sets the k-Nearest Neighbor classifier is able to predict quasars with an error of $\approx 6\%$. A cross check with all spectroscopically classified quasars determines an error of 8.62% undetected quasars.

- **Features 4:** mag$_{psf}$(u,r,g,i,z), mag$_{model}$(u,g,r,i,z)
- **Labels 4: A:** quasars, **B:** stars and galaxies

Color/SDSS Features on Selected Objects



References

[1] Nicholas M. Ball. *Data Mining and Machine Learning in Astronomy*, 2009, arXiv:0906.2173v1 [astro-ph.IM]

[2] Kirk Borne. *Scientific Data Mining in Astronomy*, 2009, arXiv:0911.0505v1 [astro-ph.IM]

[3] Sloan Digital Sky Survey. `http://www.sdss.org`, June 2010.

[4] Trevor Hastie, Robert Tibshirani und Jerome Friedman. *The Elements of Statistical Learning*, Springer, 2010.

[5] B. W. Matthews, *Comparison of the predicted and observed secondary structure of T4 phage lysozyme* ,Biochimica et Biophysica Acta, 1975.

[6] Ethem Alpaydin, *Introduction to Machine Learning*, MIT Press, 2010.

[7] The Sloan Digital Sky Survey Quasar Catalog V. Seventh Data Release. `http://arxiv.org/abs/1004.1167`, June 2010.

Acknowledgements
We thank Anna Amelung for the cartoon.

QR-Code of Online Version

Annual Meeting of the Astronomische Gesellschaft 2010, Bonn, Germany, 13-17 September 2010.