Discriminating Point and Extended Sources with k-Nearest Neighbors

Kai Lars Polsterer[1] polsterer@astro.ruhr-uni-bochum.de

Fabian Gieseke[2] fabian.gieseke@tu-dortmund.de

Andreas Thom[2] andreas.thom@tu-dortmund.de

Oliver Kramer[2] **Jan Vahrenhold**[2] **Dominik Bomans**[1] **Ralf-Jürgen Dettmar**[1]

[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

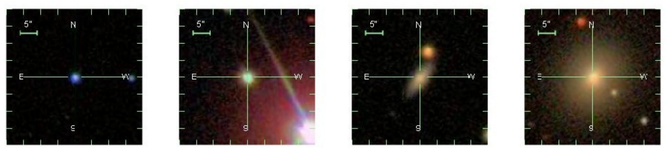
The task of discriminating point and extended sources is important for spectroscopic target selection. We present a classification approach based on k-Nearest Neighbors and evaluate its performance on a generated data set. Our experiments indicate that the classifier exhibits a superior performance compared to the standard approach used in the Sloan Digital Sky Survey (SDSS) pipeline.

Motivation

Classification Task: Point & Extended Sources

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 consider the task of discriminating point and extended sources by using precomputed features available in the SDSS.

Point & Extended Sources



Source: http://www.sdss.org

Data

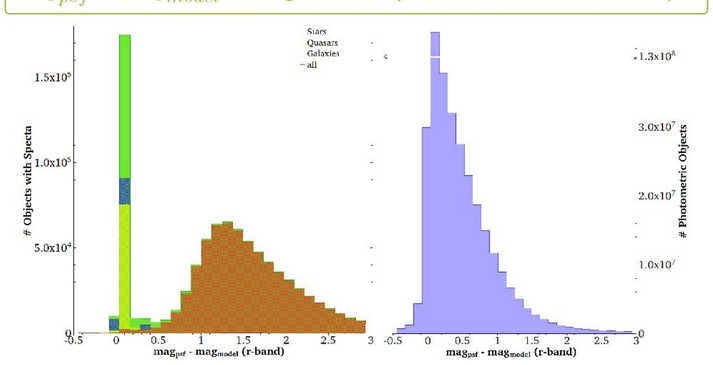
Our data set is obtained from photometric and spectroscopic data available in the Sloan Digital Sky Survey (DR6) which is said to be "one of the most ambitious and influential surveys in the history of astronomy" [3]. The data for this survey has been taken via a 2.5-meter telescope at the Apache Point 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

The data sets we use have been 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 creates two major populations. In comparison to the distribution of all photometric data sets, objects with a low spatial extension are not covered accordingly. Therefore we created two data sets. One containing 100,000 randomly selected objects and another one with 82,344 objects that have been selected to reproduce the distribution of all photometric objects.

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



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

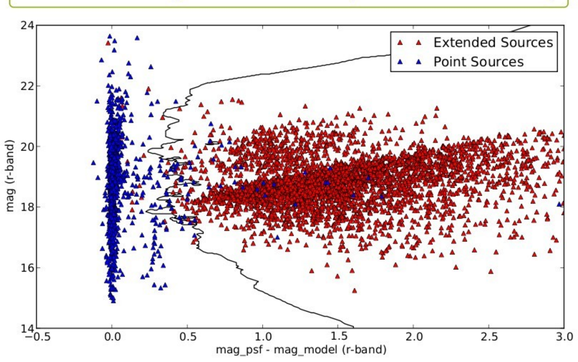
$$\hat{Y}(\mathbf{x}) = \begin{cases} 1 & \text{if } f(\mathbf{x}) > 0 \\ -1 & \text{if } f(\mathbf{x}) \leq 0, \end{cases} \quad (1)$$

where

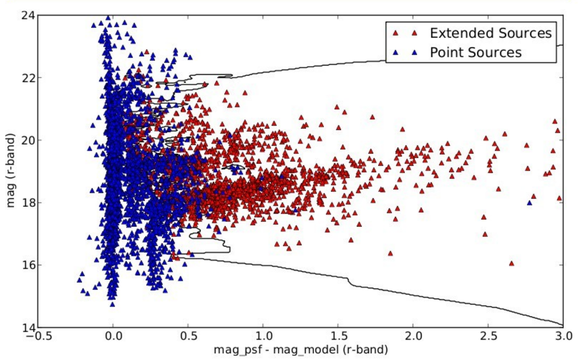
$$f(\mathbf{x}) = \frac{1}{k} \sum_{\mathbf{x}_i \in N_k(\mathbf{x})} y_i \quad (2)$$

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 (Random Data Set)



k-Nearest Neighbor Classifier (Selected Data Set)



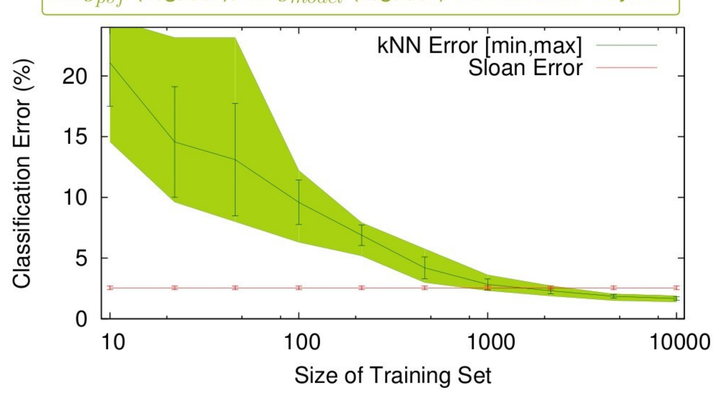
Experiments

Experimental Setup: 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. All other objects are classified as "point sources".
In the presented 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.
To analyze the performance of the classifiers an 8-fold cross-validation [4] is used. In 8 consecutive runs the data set is split into a training set (7/8) and a test set (1/8). On this 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 omitted from the plots.

Experiment 1: The SDSS classifier performs well on the randomly selected objects due to the fact that a similar classifier was used to select these objects. Only with a very high number of training sets the k-Nearest Neighbor classifier outperforms the SDSS classifier.

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

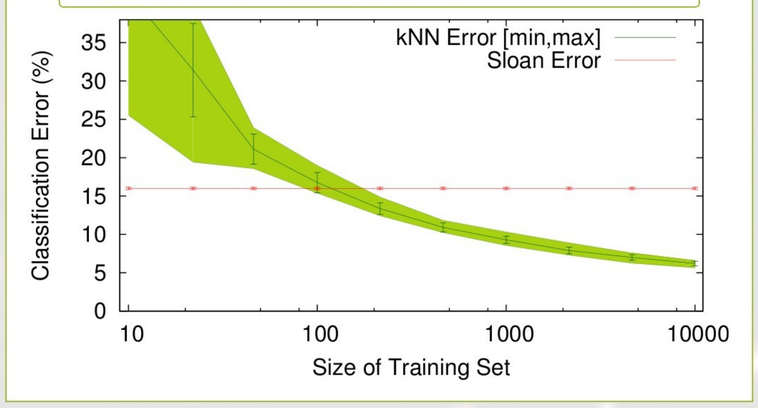
mag_{psf} (u,g,r,i,z), mag_{model} (u,g,r,i,z) on Random Objects



Experiment 2: The SDSS classifier has an error of $\approx 16\%$ on the selected data set that reproduces the distribution of all photometric objects. Equivalent results are achieved by the k-Nearest Neighbor classifier for a training set with roughly 200 elements. Given more elements the classifier significantly outperforms the SDSS classifier.

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

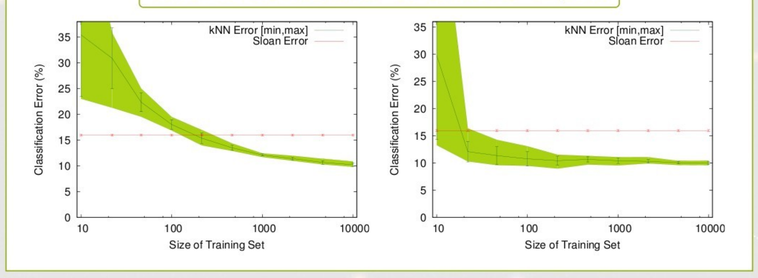
mag_{psf} (u,g,r,i,z), mag_{model} (u,g,r,i,z) on Selected Objects



Experiment 3 a,b: When using only the colors as features the performance of the classifier drops. The same effect can be observed when using the difference between the psf and model magnitudes. Both experiments have been carried out with the selected data set.

- **Color Features:** u-mag$_{psf}$ - g-mag$_{psf}$, g-mag$_{psf}$ - r-mag$_{psf}$, r-mag$_{psf}$ - i-mag$_{psf}$, i-mag$_{psf}$ - z-mag$_{psf}$
- **SDSS Features:** g-mag$_{psf}$ - g-mag$_{model}$, r-mag$_{psf}$ - r-mag$_{model}$, i-mag$_{psf}$ - i-mag$_{model}$
- **Labels: 1:** stars and quasars, **2:** galaxies

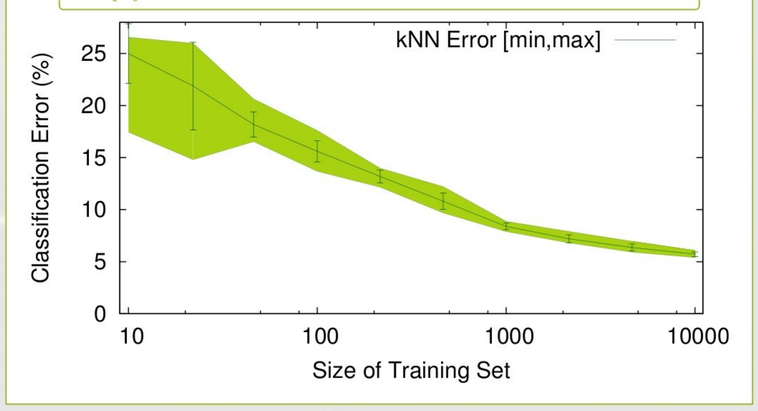
Color/SDSS Features on Selected Objects



Experiment 4: When trained with 10,000 training 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:** u-mag$_{psf}$, g-mag$_{psf}$, r-mag$_{psf}$, i-mag$_{psf}$, z-mag$_{psf}$, u-mag$_{model}$, g-mag$_{model}$, r-mag$_{model}$, i-mag$_{model}$, z-mag$_{model}$
- **Labels: 1:** quasars, **2:** stars and galaxies

mag_{psf} (u,g,r,i,z), mag_{model} (u,g,r,i,z) 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.

Acknowledgements We thank Anna Amelung for the cartoon.