DISC-Quasars: Identification of Distant Quasars in Sky Surveys

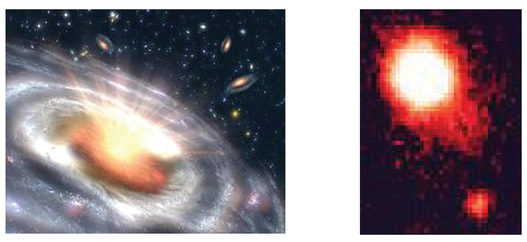
Bin Fu, Sangjae Yoo, Gor Nilanon, Eugene Fink, Julio López, and Garth Gibson

We are developing data-mining techniques for detection of distant quasars in sky-survey datasets, which will help astronomers to distinguish quasars from other celestial objects, such as stars and galaxies, based on analysis of telescope images made through different color filters.

Problem

A *quasar* is a galaxy with an unusually massive black hole in its center, which may be hundreds of times as bright as a regular galaxy. Astronomers use distant quasars as "beacons", which allow charting remote regions of the universe and studying its expansion.

Accurate identification of quasars is a hard problem, since they look like "shiny dots", not much different from stars and regular galaxies. To distinguish objects of different types, astronomers compare images made through five color filters, and study the distribution of each object's brightness over these colors. We are working on application of data mining to this problem, specifically, finding quasars that are at least 12 billion light years away.

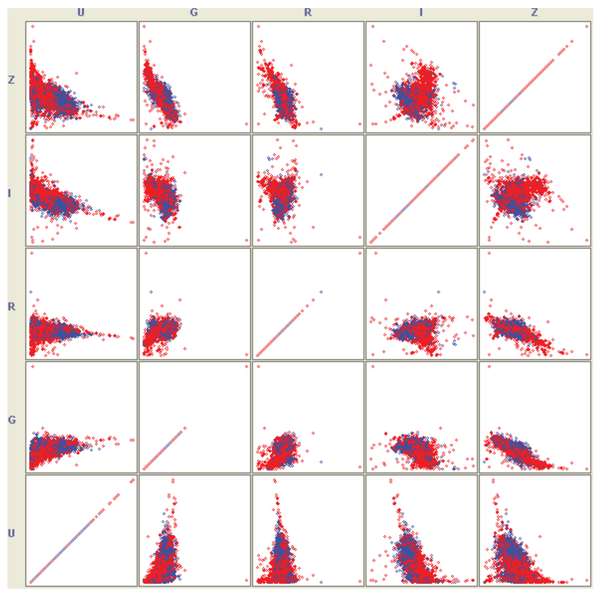


Quasars: An artist's impression (left) and a telescope image (right).

Approach

We represent a celestial object by five numeric values, which show its brightness in images made through five color filters, and apply supervised learning to identify distant quasars based on these values. We have experimented with the following learning techniques and their combinations:

- **Support vector machines**
- **Decision trees**
- **Nearest neighbors**
- **k-means clustering**



Quasars (blue) and non-quasars (red) in the space of five color brightnesses, denoted U, G, R, I, and Z.

Results

We use two performance metrics:

- **Precision**: **percentage of identified objects that are true quasars**
- **Recall**: **percentage of quasars that have been identified**

Technique	Precision	Recall
Support vector machines (with RBF kernel)	75%	55%
Decision trees (using the C4.5 algorithm)	71%	61%
Nearest neighbors (with 11 neighbors)	73%	60%
k-means clustering (with k = 14)	83%	22%
Majority-vote combination of decision trees, support vector machines, and nearest neighbors	79%	56%

Conclusions:

- ***k-means clustering***: **the highest precision at the expense of a low recall**
- ***Majority vote***: **a good combination of a high precision and a reasonable recall**

The resulting precision is twice as high as that of the earlier methods developed by astronomers.

Carnegie Mellon University

Parallel Data Laboratory
PDLR10