Mining the CSC to Identify Classes of AGNs

R. D'Abrusco et al.

Harvard-Smithsonian Center for Astrophysics

Observations of Active Galactic Nuclei (AGNs) are fundamental for modern astrophysics because they shed light on both Black Hole (BH) physics and galaxy formation and evolution, and on the role of AGN feedback in these processes. X-rays observations are efficient at selecting AGNs, but do not provide the full picture. Convincing evidences have been recently gathered pointing to several peculiar classes of AGNs: obscured quasars, intrinsically red quasars and even quasars missing their 'hot dust bump'. Different selection criteria have given different views of obscuration in quasars, and multi-dimensional analyses of multi-wavelength observations are emerging as a powerful tool in this quest. We propose a new method for AGNs classification through a Data Mining (DM) technique applied to the Chandra Source Catalogue (CSC) data, complemented by archival data at several different wavelengths. This method will allow a multi-wavelength characterization of the Spectral Energy Distribution (SED) of X-ray selected AGNs and will help to determine the correlations between sources selected with different criteria and relate them to their X-ray properties.

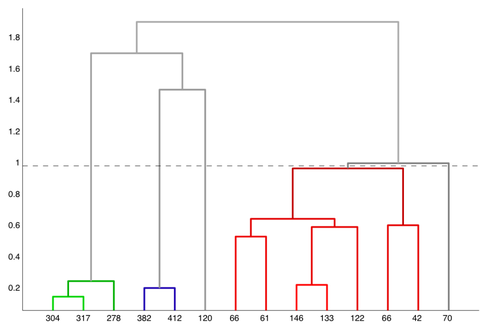
DM is now a mature tool for astronomical research thanks to the convergence of strong scientific motivations, large databases, powerful computers and the development of advanced statistical pattern recognition techniques.

Outline of the project

- We explore the distribution of AGNs in the high dimensionality parameter space obtained by combining all available multi-wavelength data in order to identify the distinctive features of the SEDs of the CSC sources through unsupervised clustering and the use of a suitable Base of Knowledge (BoK).
- Our projects involves a DM study of the CSC supplemented by the following catalogues: **SDSS** for optical, **UKIDSS** for NIR, **GALEX** for UV and **VLA-First** for radio.
- The DM approach is entirely data driven, not relying on any predetermined selection techniques derived by observations in a specific band.

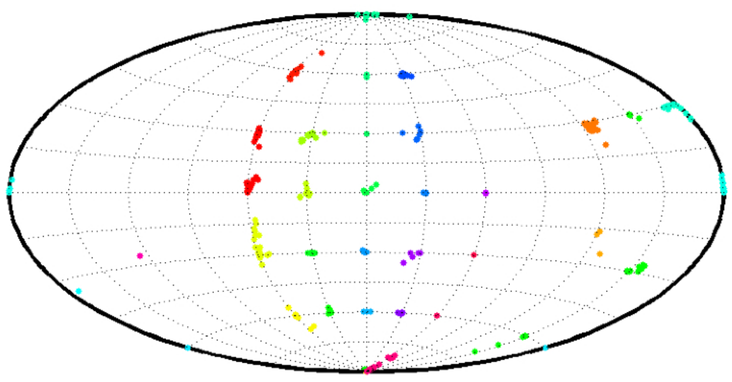
The method

We will describe statistically individual 'clusters' using an a-priori built set of examples ('base of knowledge', BoK). This method yields a clear advantage over the traditional study of low-dimensional correlations, for two main reasons:



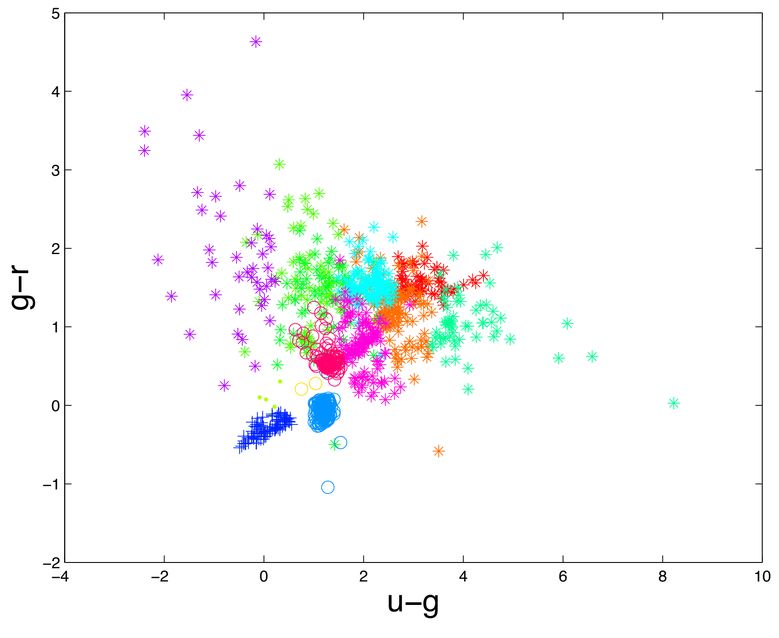
(1) it allows elusive correlations between parameters to show their signature;

(2) efficiently meshes different types of information, continuous (e.g., color, fluxes), and categorical data (e.g., classification indices, type 1, type2) (Murtagh 2003).



Our DM approach, already applied to the problem of extraction of optical candidate quasars from SDSS database (D'Abrusco et al. 2009, Laurino et al. in prep.), is based on the combination of two algorithms:

- **Probabilistic Principal Surfaces** (PPS, Chang & Ghosh 2000): a generalization of Principal Component Analysis, it is used to



perform unsupervised clustering throug dimensionality reduction;

- **Negative Entropy Clustering** (NEC): a hierarchical clustering algorithm based on the Negentropy, i.e. the difference between the information needed for describing a random distribution and that needed to describe a given multidimensional distribution.

Although they identify new and 'spontaneous' associations of sources, clustering algorithms are not predictive. Additional information, i.e the BoK, can be used to label the points of the clustered distribution. Labeled points are then used as tracers of particular clusters and can be used to extract new members of that family and isolate clusters of objects not present in the BoK and, therefore, potentially interesting.

The statistical challenge

Two different aspects of the application of this clustering technique to our dataset have to be considered:

- **Censored analysis**: detections may be missing in some of the wavelengths (e.g. a CSC source may have SDSS but not UKIDSS counterpart). We plan to introduce censoring in the DM, by using the sensitivity information o f each catalog to estimate limits and the BoK to label limits as such. With this approach, it will be possible to characterize sources within the CSC fields but with no X-ray detection. The validity of this method will be assessed through simulation;
- **'Curse of dimensionality'**: volume grows exponentially with the number of dimensions, so that the distribution of sources becomes sparser and sparser as new observed quantities are added to the features space.

Expected results

The primary purpose of this study is to provide new observational constraints to the understanding of AGNs and their evolution by obtaining as wide as possible a census of AGN behavior in the 13-dimensional space of *XUV1UV2ugrizYJHKRad* and by identifying possible new peculiar rare classes of AGNs. The final outcome of the CSC mining will be an objective classification of all CSC sources.

CSC

The CSC (Evans, I. N., et al. 2010, ApJS, in press) is an X-ray virtual facility providing access to a large set of quantities measured for ~10^5 point and compact sources (spatial extension < 30"), detected in a subset of public ACIS imaging observations of the Chandra mission. The catalogue, exposed both as a web application and a web service, provides a unique tool to 1) retrieve reliable measurements of observable quantities; 2) to facilitate the statistical analysis of several properties of a reliably selected sample of X-ray source; 3) give access to calibrated observational data.

CSC can be found at the URL: http://cxc.harvard.edu/csc