Highlights of Astronomy, Volume 16
XXVIIIth IAU General Assembly, August 2012
T. Montmerle, ed.

doi:10.1017/S1743921314012885

Knowledge Discovery Workflows in the Exploration of Complex Astronomical Datasets

Raffaele D'Abrusco[1], Giuseppina Fabbiano[1], Omar Laurino[1] and Francesco Massaro[2]

[1]Harvard-Smithsonian Center for Astrophysics - Cambridge (MA), 02138 - Garden Street 60
[2]SLAC National Laboratory and Kavli Institute for Particle Astrophysics and Cosmology, 2575 Sand Hill Road, Menlo Park, CA 94025, USA

Abstract. The massive amount of data produced by the recent multi-wavelength large-area surveys has spurred the growth of unprecedentedly massive and complex astronomical datasets that are proving the traditional data analysis techniques more and more inadequate. Knowledge discovery techniques, while relatively new to astronomy, have been successfully applied in several other quantitative disciplines for the determination of patterns in extremely complex datasets. The concerted use of different unsupervised and supervised machine learning techniques, in particular, can be a powerful approach to answer specific questions involving high-dimensional datasets and degenerate observables. In this paper I will present CLaSPS, a data-driven methodology for the discovery of patterns in high-dimensional astronomical datasets based on the combination of clustering techniques and pattern recognition algorithms. I shall also describe the result of the application of CLaSPS to a sample of a peculiar class of AGNs, the blazars.

Keywords. surveys, data-mining, AGNs, blazars, WISE, γ-ray

1. CLaSPS

The Clustering-Labels-*Score* Pattern Spotter (CLaSPS) method (D'Abrusco *et al.* (2012)) for the discovery of patterns in complex astronomical *feature* spaces is based on unsupervised clustering techniques (Hastie *et al.* (2009)), complemented by additional data, the *labels*, employed to characterize the content of different clusters. The *labels* are used to characterize the content of the set of the clusters determined in the *feature* space. Previously, some of the authors ((D'Abrusco *et al.* (2009)) and (Laurino *et al.* (2011))) have used the same approach for the selection of optical candidate quasars from photometric datasets and the determination of photometric redshifts employing. CLaSPS has generalized this method extended to multiple labels, both numerical and categorial. The originality of CLaSPS lies in the criterion used to select the interesting aggregation of sources in the *feature* space that show correlations with the *labels* distribution. The quantitative diagnostics used to select such clusterings is called the *score*. For a generic clustering of the dataset in the *feature* space, the *score* of the i-th cluster is defined as:

$$S_i = \sum_{j=1}^{M^{(L)}-1} \|f_{ij} - f_{i(j-1)}\| \tag{1.1}$$

where f_{ij} is the fraction of the i-th cluster members with values of the *label* falling in the j-th bin (or class) and the sum is over all *labels* bins. The total *score* s of the clustering follows:

$$S_{\rm tot} = \frac{1}{N_{\rm clust}} \cdot \sum_{i=1}^{N_{\rm clust}} S_i = \frac{1}{N_{\rm clust}} \sum_{i=1}^{N_{\rm clust}} \Big(\sum_{j=1}^{M^{(j)}-1} \| f_{ij} - f_{i(j+1)} \| \Big) \tag{1.2}$$

The values of the *scores* are then used to select the clustering(s) showing the largest degree of correlation between the *label* classes and clustering membership distributions (left plot in Fig. 1). The effectiveness of the *score* has been assessed on simulated clusterings before the application to real astronomical datasets.

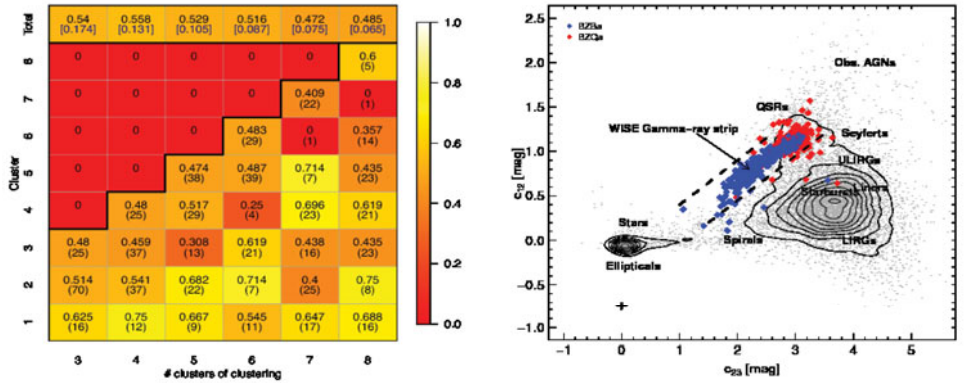

Figure 1. Left: map of the *scores* for the clusterings of the blazars experiment described in Sec. 2 as a function of the total number of clusters in each clustering. Right: projection of the blazars WISE *locus*, discovered by CLaSPS, onto the WISE [4.6]−[12] vs [3.4]−[4.6] color-color plane.

2. Application to Blazars

One of the first applications of CLaSPS, involving a sample of *bona fide* blazars in a *feature* space generated by broad-band color from mid-infrared to far ultraviolet, has led to the discovery of a previously unknown pattern followed by the blazars in the mid-infrared color space generated by the WISE magnitudes (D'Abrusco *et al.* (2012)). CLaSPS has been applied to the distribution of blazars in the 9-dimensional colors *feature* space using as *labels*, among other observables, the blazars spectral classification in BL Lacs and Flat Spectrum Radio Quasars, and the detection of γ-ray emission. CLaSPS determined a significant pattern for γ-ray emitting blazars of both spectral types in the three dimensional WISE colors space, revealing that this class of extragalactic sources occupy a peculiar and narrow *locus* in this *feature* space. The projection of the 3D WISE blazars *locus* color plane is shown in the right plot in Fig. 1. This discovery has also been used to devise a method for the selection of WISE candidate blazars that has been already applied to different sample of high-energy unidentified sources (the application to the unidentified γ-ray sources from Fermi in (Massaro *et al.* (2012)).

References

D'Abrusco, R., Longo, G., & Walton, N. A. 2009, *MNRAS*, 396, 223
D'Abrusco, R., Massaro, F., Ajello, M., *et al.* 2012, *ApJ*, 748, 68
D'Abrusco, R., Fabbiano, G., Djorgovski, G., *et al.* 2012, *ApJ*, 755, 92
Hastie, T., Tibshirani, R., & Friedman, J. 2009, *The Elements of statistical learning*, Springer.
Laurino, O., D'Abrusco, R., Longo, G., & Riccio, G. 2011, *MNRAS*, 418, 2165
Massaro, F., D'Abrusco, R., Tosti, G., *et al.* 2012, *ApJ*, 750, 138