The exploration of multi-wavelength astronomical datasets: the case of AGNS in the Chandra Source Catalog and unsupervised clustering.

R. D’Abrusco
Motivations

Characterization of the distribution of AGNs in a high dimensionality parameter space obtained by combining multi-wavelength data and study of their X-rays properties.

The primary purpose of this study is to obtain a possible census of AGN behavior in the 13-dimensional features space of X-UV-optical-IR-Radio photometry, and pick up outliers and constrain their nature.
The approach

Unsupervised clustering in a high dimensional features space

AND

use additional information (labels) to identify interesting cluster(ing)s

IN ORDER TO

derive new high dimensional correlations and/or expand known correlations to other bands, and/or spot unusual behaviors (outliers).
Statistical issues

Few points for a large space

> 10 dimensional features space
$10^2 \sim 10^3$ sources

Upper limits & Clustering

Inclusion of sources with no detections but reliable upper limits.

Outliers vs Clusters

Most clustering methods tends to prefer the selection of either well populated homogeneous clusters or to sparse clusters/singletons (outliers).
The datasets

1) “Large area surveys” sample

SDSS quasars

SDSS  \( (9262) \)  UKIDSS  \( (7405) \)  GALEX  \( (112) \)  CSC  \( (195) \)  CSC

2) Chandra COSMOS X-ray survey

3) SWIRE
The datasets

1) “Large area surveys” sample

Survey databases with crossmatched catalogs, TAP

2) Chandra COSMOS X-ray survey

3) SWIRE
Multiple techniques

Hierarchical Clustering (K-means)

Self-Organizing Maps (SOM)

Principal Probabilistic Surfaces
Multiple techniques

Hierarchical Clustering (K-means)

Self-Organizing Maps (SOM)

Principal Probabilistic Surfaces
Dendrograms

Representation of hierarchical structure - HC tree.

HC does not require a fixed number of clusters and produces all possible clustering, given a measure of dissimilarity (distance). Every generation of clusters maximizes the between-group dissimilarity.

dissimilarity ≡ (metric, linkage strategy)

Metrics:
- Euclidean, Manhattan,
- Mahalanobis, maximum,
- etc.

Linkage strategies:
- complete, single,
- average, etc.
Distance: euclidean; linking strategy: complete
Distance: euclidean; linking strategy: complete
Distance: euclidean; linking strategy: complete
Labels to pick clusters

Other measured quantities, called “labels” (continuous or discrete) are separated into bins (if continuous) and used to pick those cluster(ing)s which seem interesting.

$L_X$, $HR$, $\Gamma$, $n_h$, X-ray/optical ratio, X-ray time variability, radio morphology (if available)
Discriminating clusterings

The score, a modified version of the "total variation" of a feature vector:

\[
S_{TOT} = \sum_{i=1}^{N_{Cl}} S_i = \frac{\sum_{i=1}^{N_{Cl}} \left( \sum_{j=1}^{K-1} ||f_{ij} - f_{i(j+1)}|| \right)}{N_{Cl}}
\]

Histograms of scores:

- random
- partly correlated
- correlated
### Score distribution for HC clustering, relative to HR.ms. label

<table>
<thead>
<tr>
<th>Cluster index</th>
<th>Total</th>
<th>0.573</th>
<th>0.547</th>
<th>0.564</th>
<th>0.609</th>
<th>0.596</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>0.524</td>
<td>0.547</td>
<td>0.596</td>
<td>0.609</td>
<td>0.573</td>
<td>0.580</td>
</tr>
</tbody>
</table>

### Score distribution for HC clustering, relative to HR.hs. label

<table>
<thead>
<tr>
<th>Cluster index</th>
<th>Total</th>
<th>0.306</th>
<th>0.356</th>
<th>0.425</th>
<th>0.46</th>
<th>0.453</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>0.307</td>
<td>0.306</td>
<td>0.356</td>
<td>0.425</td>
<td>0.460</td>
<td>0.453</td>
</tr>
</tbody>
</table>

KISS Study Program - 6/9/11
<table>
<thead>
<tr>
<th>Label name</th>
<th>Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>L.b.</td>
<td>0.492</td>
</tr>
<tr>
<td>HR.hs.</td>
<td>0.374</td>
</tr>
<tr>
<td>HR.ms.</td>
<td>0.507</td>
</tr>
<tr>
<td>z</td>
<td>0.608</td>
</tr>
</tbody>
</table>
Experts for $z_{\text{phot}}$

Fuzzy k-means

Cluster 3
Cluster 5
Cluster 2
Cluster 1
Cluster 6
Cluster 7
Cluster 4
Cluster 8

Gating Network

Experts

$z_{(1)}_{\text{phot}}$, $z_{(2)}_{\text{phot}}$, ..., $z_{(N)}_{\text{phot}}$

$z_{(\text{Best})_{\text{phot}}}$

KISS Study Program - 6/9/11
Experts for other problems

Fuzzy k-means

Gating Network

Cluster 8
Cluster 3
Cluster 5
Cluster 1
Cluster 2
Cluster 4
Cluster 6
Cluster 7
Cluster 1
Cluster 2
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8

Experts for other problems

KISS Study Program - 6/9/11
Conclusions

Results for the first dataset and Chandra COSMOS survey with HC

Open issues:

- clustering with upper limits/no detection/missing data (model based? simulations?)

- labels “binning”: is there a data-driven way to accomplish this? Co-clustering is being explored.

Integration of different methods in a Gated Expert model:

- example 1: template fitting and machine learning experts for $z_{\text{phot}}$ calculation.

- example 2: extraction of candidate quasars from optical photometric quasars.