Learn from every mistake!

Hierarchical information combination

Maria Süveges

Joint work with S. Fotopoulou, J. Coupon, S. Paltani, L. Rimoldini, L. Eyer

AstroInformatics 2016, October 20-24, Sorrento, Italy
A fundamental feature:
many methods, different performances
A fundamental feature:
many methods, different performances

Method 1
Method 2
Method 3
Method n
A fundamental feature: many methods, different performances

HOW TO COMBINE THEIR STRENGTHS?

My examples:

• photometric redshift estimation
  photometry $\rightarrow Z_{\text{phot}}$

• variable star classification
  period, amplitude, Fourier parameters, colour $\ldots \rightarrow$ variability class
Variable star classification (Hipparcos, OGLE, Gaia, Pan-STARRS, LSST, …)

Data: Hipparcos periodic variables
(European Space Agency, 1997, The Hipparcos and Tycho catalogues (ESA SP-1200))

Global accuracies:

- RF: 82.5%
- C5.0: 81.0%
- GM: 79.9%
- SVM: 79.9%
- LDA: 77.9%
Photometric redshift estimation (Euclid, Pan-STARRS, LSST, …)

Data:
photometry: WIRDS+CFHTLS (Bielby et al., 2012, A&A 545, A23)
Photometric redshift estimation (Euclid, Pan-STARRS, LSST, …)

Data:
photometry: WIRDS+CFHTLS (Bielby et al., 2012, A&A 545, A23)
Combination of the output of all: a potentially good solution...

- Training set
- Combination set
- Validation set

Method 1 train
Method 2 train
Method 3 train
Method 4 train
Method 5 train

\[ p_{RF}^{1, C_K}, \ldots, p_{LD}^{1, C_K}, \ldots, p_{RF}^{2, C_K}, \ldots, p_{LD}^{2, C_K} \]

Various averages

\[ p_{comb}^{1, C_1}, p_{comb}^{1, C_2}, \ldots, p_{comb}^{2, C_1}, p_{comb}^{2, C_2}, \ldots \]
...but limitations of combination by average: a restricted domain
...but limitations of combination by average: a restricted domain
...but limitations of combination by average: a restricted domain
...but limitations of combination by average: a restricted domain

We need the mapping (Method 1, Method 2) \(\rightarrow\) (hopefully) best achievable

This mapping is not necessarily linear!
There is very little we know a priori about it (dependence between base methods, their relation with truth, ...)
A potentially good solution: learn a nonlinear mapping.
Classification of variable stars: general improvement

Global accuracies:
- Hier.: 85.8%
- W.avg.: 82.2%
- RF: 82.5%
- C5.0: 81.0%
- GM: 79.9%
- SVM: 79.9%
- LDA: 77.9%
Photometric redshifts: general improvement

Def. of outlier: \[
\left| \frac{z_{\text{spec}} - z_{\text{phot}}}{1 + z_{\text{spec}}} \right| > 0.15
\]

\[
\sigma_{\text{NMAD}} = 1.48 \times \text{median} \left( \left| \frac{z_{\text{spec}} - z_{\text{phot}}}{1 + z_{\text{spec}}} \right| \right)
\]
Photometric redshifts:
what linear combinations cannot give

Template fit

Empirical (Random Forest)

Combination
Summary

The method is...

- **improving** on single-method analysis: is better than !
- capable of **bias correction** (in case the training set contains relevant information)
- **general**: applicable for regression (photo-z) and for classification (variable stars)
- **simple** and **easy to tune**

Open questions:

- Test in Euclid photo-z data challenge and in Gaia variability data
- Choice of combiner? Iterative application (more than one combiner, several levels)...? Use as element in a more complex procedure (cf. Cavouti et al., Estevez et al., Tuccillo et al., this conference)
- Theoretical questions (convergence, extracted information, …)?

**It might be better to try to invent just a very original method than the best!**
Thank you!