“Machine-learned Identification of RR Lyrae Stars from Sparse, Multi-band Data: The PS1 Sample”
Sesar et al. 2017, AJ, 153, 204

GAゼミ@170607　担当：田中幹人

【アブスト】

PanSTARRS1 3πサーベイの天体からRR Lyareを選び出す方法を開発した．先行研究（Hernitschek+16）のランダムフォレスト法より，勾配ブースティング木で学習させた方がRR Lyraeの検出率＆純度ともに向上した．距離としては120kpcくらいまでのRR Lyrae星が検出可能で，80kpcの地点で，completeness 80%，purity 90%の精度である．DESやLSSTのデータにも適用可能で，LSSTを使えば，より遠方，そしてより銀河面（減光により暗い）におけるRR Lyare探査が可能となる．Gaia WEAVEやDESI(2018年にスタートする4m望遠鏡でFoVが3度の分光器)で1度あたりPS1のRRLが1個は入るので大規模サーベイでのあまりファイバーを使った抱き合わせ観測にちょうど良い．

【ゴール】

銀河系古成分の構造などの理解．

【オリジナリティ】

分類のアルゴリズムをランダムフォレストから勾配ブースティング木に変えたところ．

【ロジック】

Training Set

PS1データ（5億天体）

SDSS Stripe 82，Dracoのカタログとマッチした，190万天体のPS1データを用意．RRLが既知のSDSSなどのデータで，type ab/c RR Lyraeおよびnon RRLの3つでラベル化

classified
PS1 RRLデータ（24万天体）

・暗くなると精度が悪くなる．
・RRcよりRRabの方が精度が良い．
・さらに絞った天体45,000個の空間分布を見ると，Sagittariusストリーム(leading)のや，UMi, Dra, Sex, Sgrの4つの矮小銀河も見える．しかし，E(B-V)=1超える場所には全然見えない．

Supervised Learning（教師あり学習）
XGBoost（勾配ブースティング木）
→n個の決定木を使って，精度に応じた重み付けをし，統合してアンサンブル学習をする（たぶん）．

学習③のマシンコストが膨大なので，190万天体を全部で学習するのではなく，学習①→学習②→とある程度数を絞りながら学習している．

学習①カラー情報
PS1 color (g, r, i, z, y)
WISE color (W1, W2)
（WISE: QSOの中間赤外線観測）

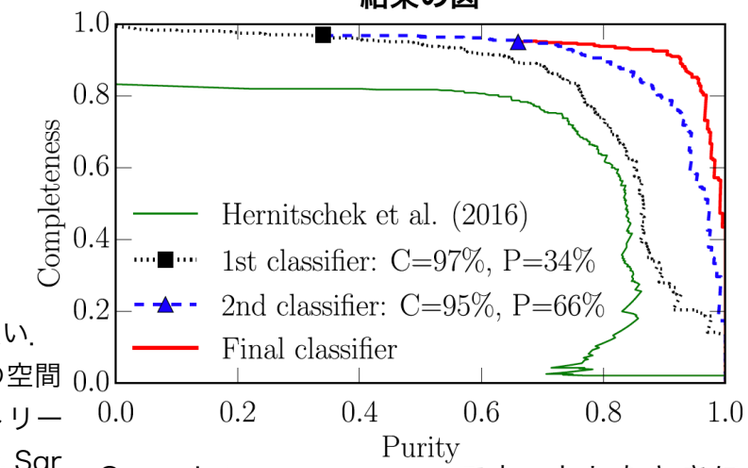
1568天体

学習②Power値
（周期解析より）

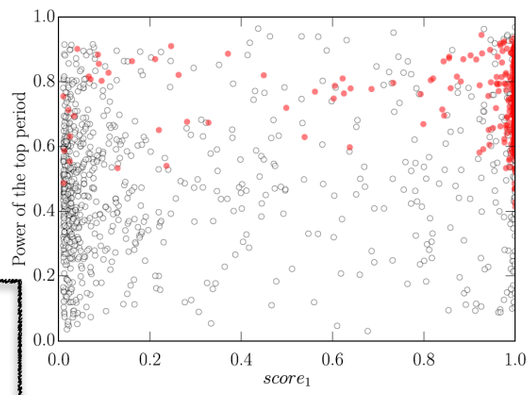
910天体

学習②multi-band光度曲線フィットの結果
（Period, χ^2, χ^2_{dof}, 尤度, 尤度$_{dof}$, J index, エントロピー S, ΔP, phase=0でのΔcolorとモデルcolor, type ab/c RRL）

結果の図

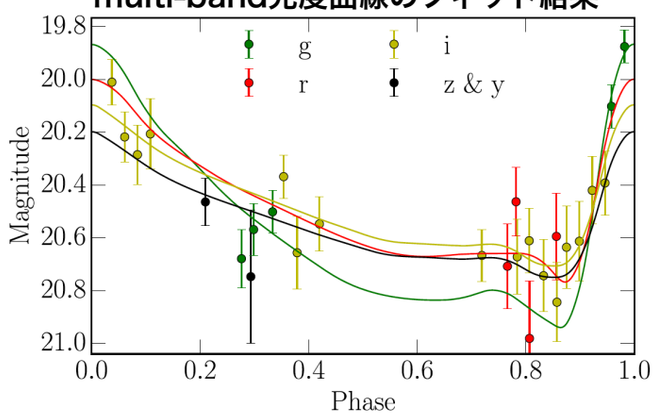


Completeness: scoreでカットしたときにrecoverされる割合．
Purity: 選んだサンプル内に本物のRRLの割合．
■→検出はされるけどコンタミが多い．
高purity→コンタミは省けるけど，本物のRRLもはじいてしまっている．

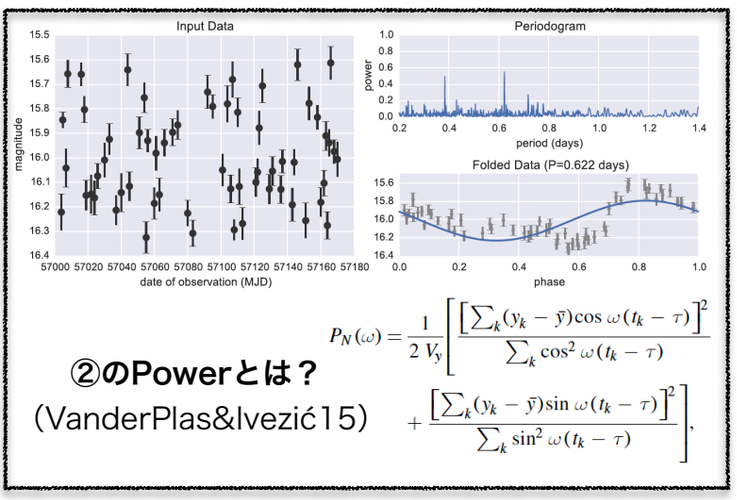


●：RR Layrae
○：non RR Layrae
※power > 0.4でもcompletenessは良好．
※scoreの添え字は各学習過程に対応．ある天体を何種類かの決定木にinputしたとき，RRLと分類される確率がscore．

multi-band光度曲線のフィット結果



pulsation model (Marconi+06)を使う



②のPowerとは？
（VanderPlas&Ivezić15）

$$P_N(\omega) = \frac{1}{2\,V_y}\left[\frac{\left[\sum_k (y_k - \bar{y})\cos\omega(t_k-\tau)\right]^2}{\sum_k \cos^2\omega(t_k-\tau)} + \frac{\left[\sum_k (y_k - \bar{y})\sin\omega(t_k-\tau)\right]^2}{\sum_k \sin^2\omega(t_k-\tau)}\right],$$

【コメント】

・大規模撮像分光サーベイに向けた，データ分析手法を洗練させる準備は驚異的．手法の開発で天体力学が発展したラプラス時代を思い起こさせる．
・試しに赤色巨星と主系列星の分類に適用してみた．Segueのデータ内で完結させるなら分類結果は良好．HSCに適用するには，トレーニングデータとしてHSCのデータとSegueのデータをマッチさせたカタログを作らないとダメそう．
→ https://www.astr.tohoku.ac.jp/~mikito/jupyter/RGBvsMS.html

Ivezić, Connolly, Vanderplas, Gray. 2014より

「Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data」

TABLE 9.1.
Summary of the practical properties of different classifiers.

Method	Accuracy	Interpretability	Simplicity	Speed
Naive Bayes classifier	L	H	H	H
Mixture Bayes classifier	M	H	H	M
Kernel discriminant analysis	H	H	H	M
Neural networks	H	L	L	M
Logistic regression	L	M	H	M
Support vector machines: linear	L	M	M	M
Support vector machines: kernelized	H	L	L	L
K-nearest-neighbor	H	H	H	M
Decision trees	M	H	H	M
Random forests	H	M	M	M
Boosting	H	L	L	L

分類方法とそのメリットデメリット

決定木の例

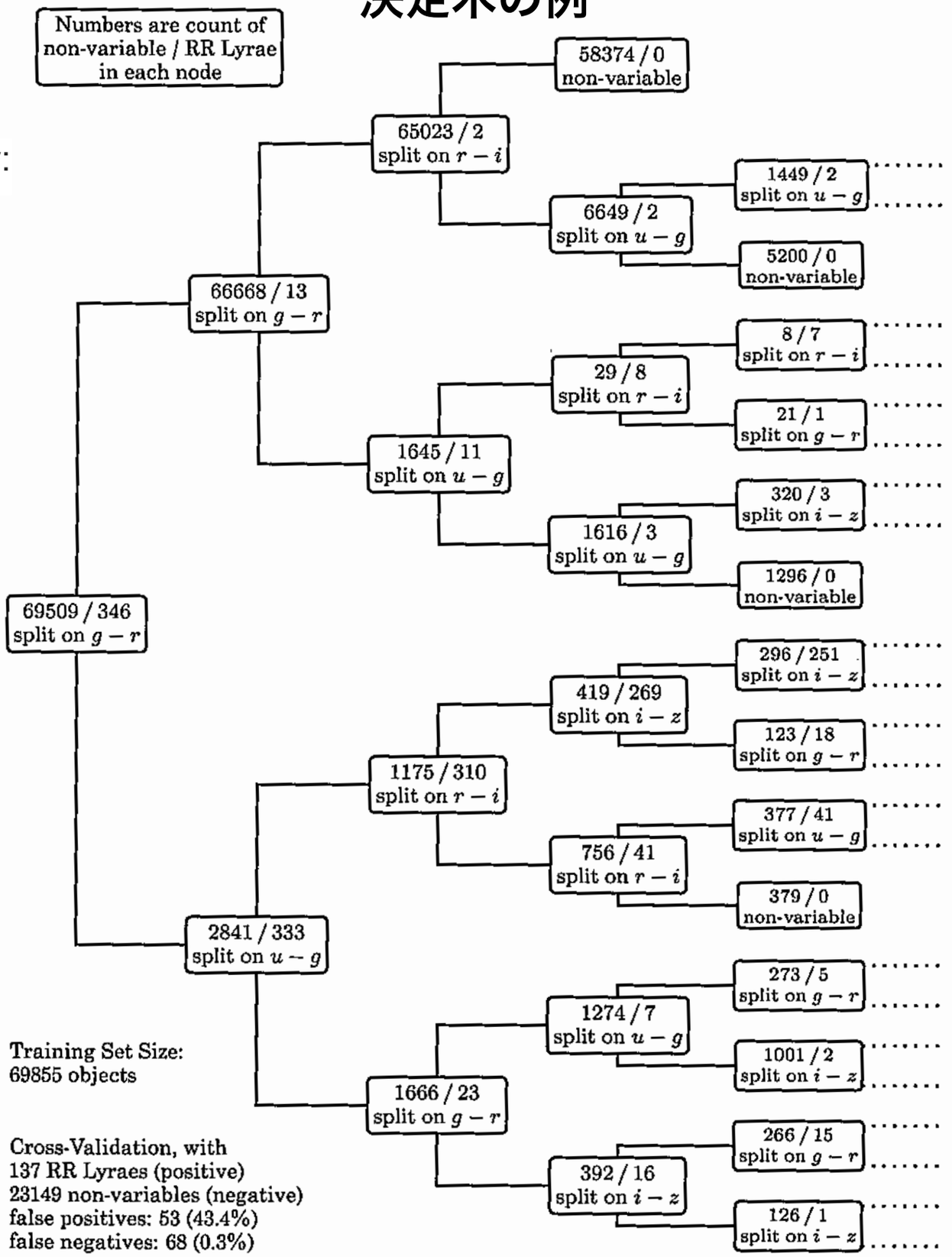


Figure 9.12. The decision tree for RR Lyrae classification. The numbers in each node are the statistics of the *training* sample of ~70,000 objects. The cross-validation statistics are shown in the bottom-left corner of the figure. See also figure 9.13.