STATISTICS
TEXAS A&M UNIVERSITY

Automated Variable Source Classification: Methods and Challenges

James Long

Texas A&M Department of Statistics

September 14, 2016

STATISTICS
TEXAS A&M UNIVERSITY

Outline

Methodology: Statistical Classifiers

Methodology: CART Example with OGLE Data

Challenge 1: Selection of Training Data

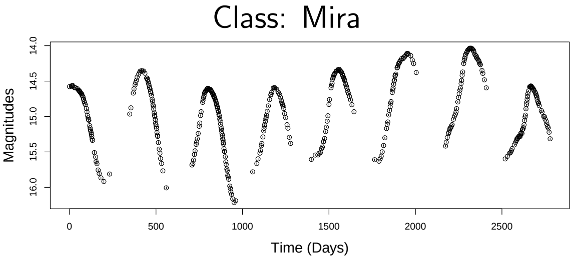
Challenge 2: Classification versus Clustering

Conclusions and Opportunities

Survey Data Sets are Large and Growing

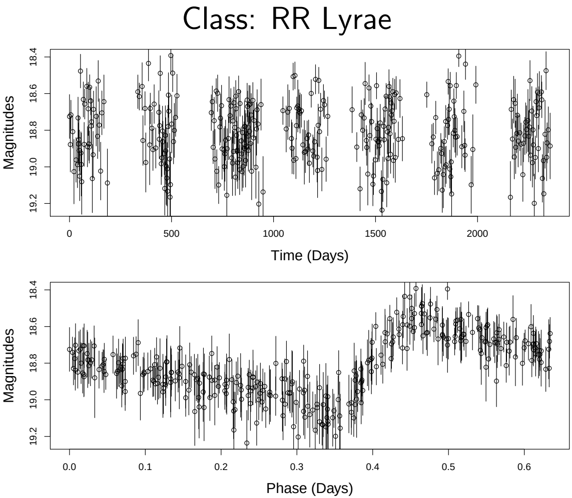
- *Hipparcos* (1989–1993): 2712 periodic variables
  - Laurent Eyer and students classified all by eye.
- *OGLE* (1992–present): 100,000s
- *Gaia* (present): millions
- *LSST* (2020): billions

Light Curves Belong to Different Classes



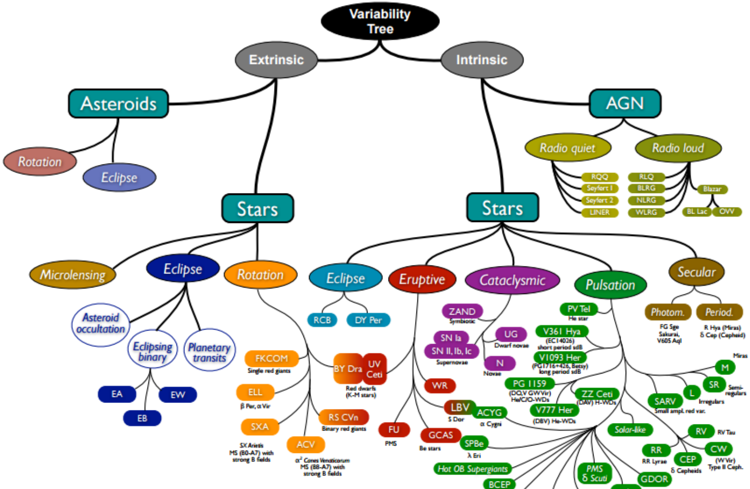
- pulsating red giant in late stage of stellar evolution
- mean magnitude variation due to dust
- long period, high amplitude

Light Curves Belong to Different Classes



- pulsating horizontal branch stars
- (almost) strictly periodic, short period, low amplitude

Hierarchical Structure to Classes

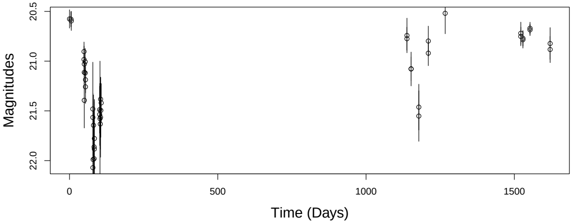


Credit : L. Eyer & N. Mowlavi (03/2009)
(updated 04/2013)

Classification Example

Data:

- $\approx 100,000$ variable sources in M33
- ≈ 30 observations / source in I–band
- mix of Miras (O–rich/C–rich), SRVs, Cepheids, non–periodic sources, junk, etc.



Goals:

- Find O–rich and C–rich Miras.
- Determine period–luminosity relationships for the Miras.

Overview of Statistical Classification

Key Terms:

- **training data:** lightcurves of known class
- **unlabeled data:** lightcurves of unknown class

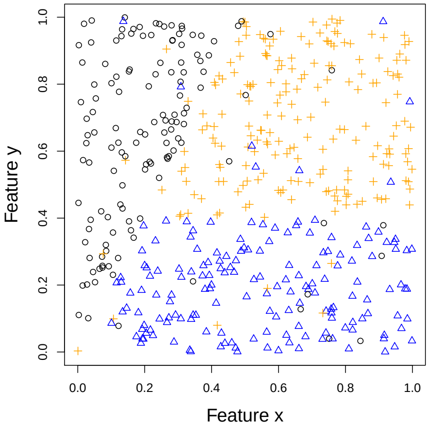
Steps in Classification:

1. **feature extraction:** derive quantities from light curves useful for separating classes, eg period, amplitude, derivatives, etc.
2. **classifier construction:** using training data, construct function

$$\widehat{\mathcal{C}}(\mathit{features}) \rightarrow \mathit{class}$$

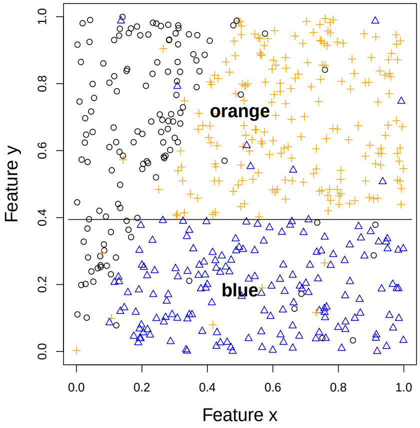
3. **apply classifier:** for unlabeled data, compute features and predict class using $\widehat{\mathcal{C}}$

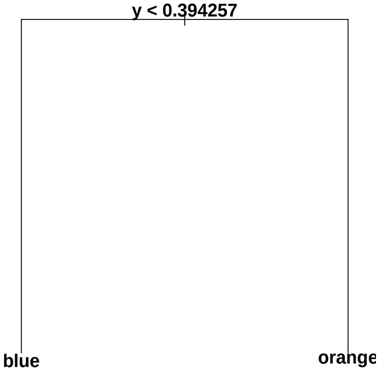
Classifier Construction using CART



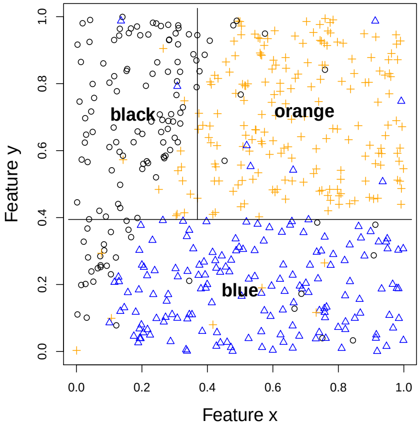
- CART developed by Breiman et al in 1980's [1]
- recursively partitions feature space
- partition represented by tree

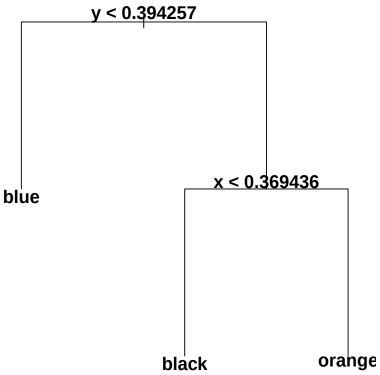
Building CART Tree . . .



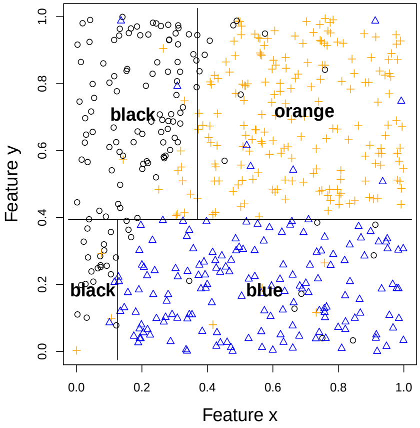


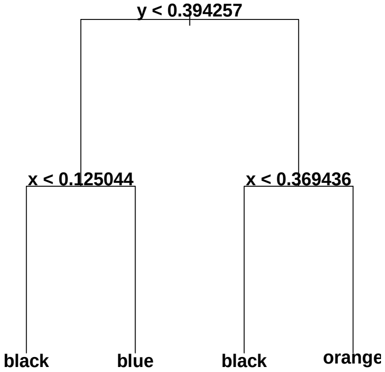
Building CART Tree . . .



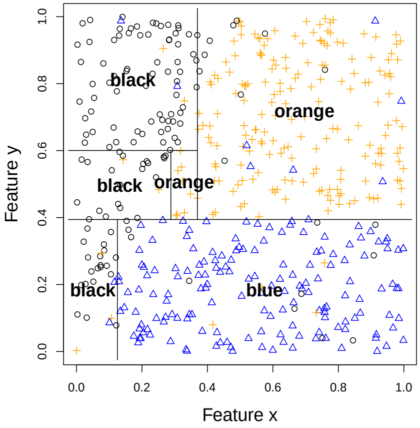


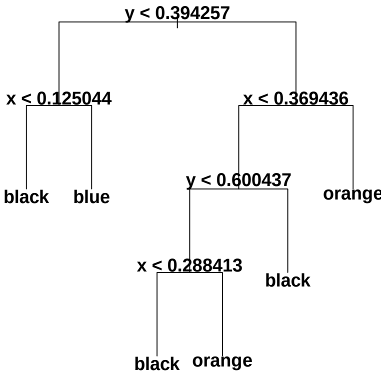
Building CART Tree . . .





Resulting Classifier





Apply Classifier to Test Data

Test Data: Data used to evaluate classifier accuracy. Test data is not used to construct classifier.

Confusion Matrix: Rows are true class of test data. Columns are predicted class of test data. Entries are counts.

	Predicted		
Truth	black	blue	orange
black	23	1	7
blue	2	30	2
orange	3	1	31

Outline

Methodology: Statistical Classifiers

Methodology: CART Example with OGLE Data

Challenge 1: Selection of Training Data

Challenge 2: Classification versus Clustering

Conclusions and Opportunities

OGLE Classification Example

Classes

- Mira O–rich
- Mira C–rich
- Cepheid
- RR Lyrae AB
- RR Lyrae C

Features

- period (of best fitting sinusoid)
- amplitude $= 95^{th}$ percentile mag - 5^{th} percentile mag
- skew of magnitude measurements
- p2p_scatter (used by Dubath)

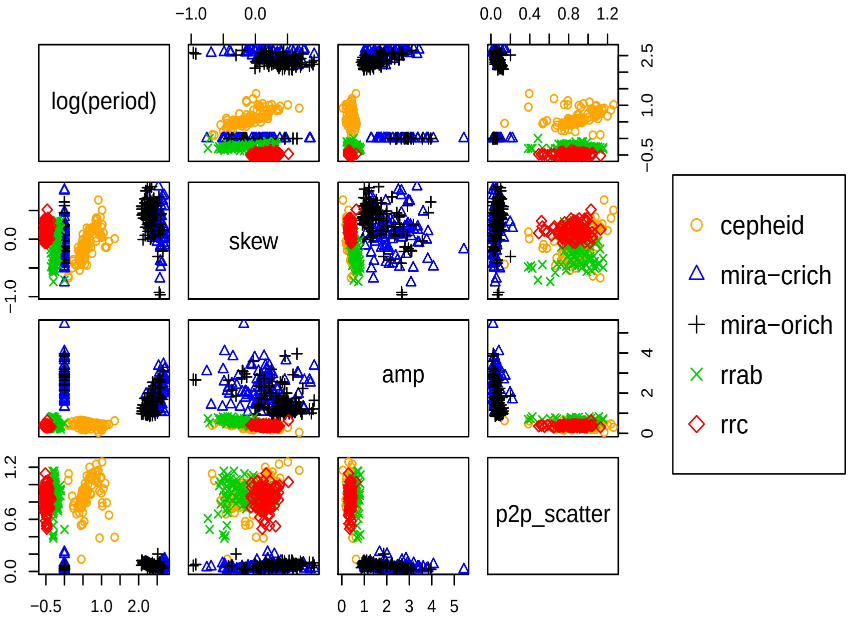
First 6 Rows of Feature–Class Dataframe

period	skew	amp	p2p_scatter	class
1.6128497	-0.5009063	0.56050	0.8672024	cepheid
0.6394983	0.3022388	0.35675	0.7523166	rrab
0.6433533	0.3200730	0.33730	0.8554517	rrab
0.4954661	-0.2053132	0.42000	0.7560226	rrab
0.3540801	0.1361693	0.34340	0.9215426	rrc
0.5460332	-0.3863142	0.69600	1.0682803	rrab

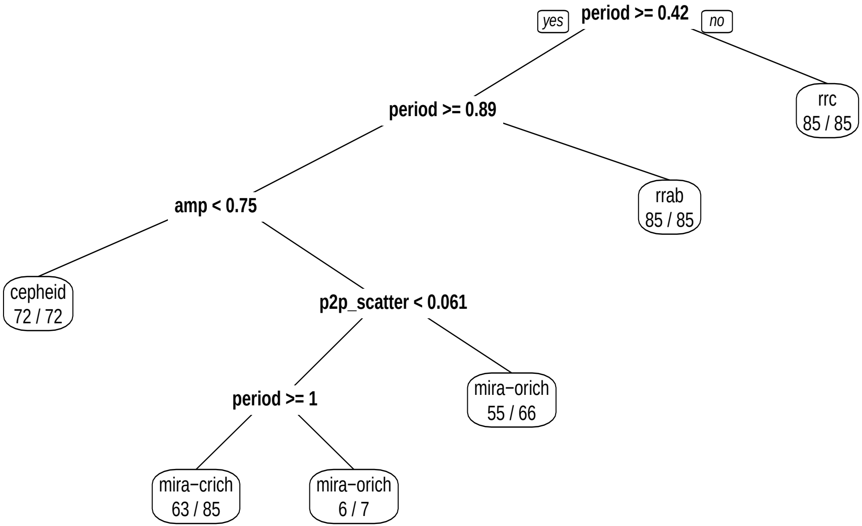
500 total rows. 5 classes.

training data: 400 randomly selected rows
test data: remaining 100 rows

Feature Distributions



CART Model Fit To Training Data



Confusion Matrix using Test Data

	Predicted				
Truth	cepheid	mira-crich	mira-orich	rrab	rrc
cepheid	24	0	0	0	0
mira-crich	0	15	10	0	0
mira-orich	0	5	12	0	0
rrab	1	0	0	14	0
rrc	0	0	0	1	14

Conclusion: Develop features to better separate O/C–rich Mira.

Notes on Existing Classification Literature

- **"On machine learning classification of variable stars" Richards J. et al. 2011 [7]**
  - mix of OGLE and Hipparcos data
  - extract 50+ features
  - test several classifiers, Random Forest works best
- **"Random forest automated supervised classification of Hipparcos periodic variable stars" Dubath et al. 2011 [2]**
  - Hipparcos data
  - extract $\sim$ 10 features
  - use random Forest
- **"Modeling Light Curves for Improved Classification" Faraway, J. Mahabal, A. et al. 2014 [3]**
  - model light curve variation using Gaussian processes
  - extract features from Gaussian process fit
  - improve classification accuracy over simpler features used in [7]

Outline

What Training Data To Use?

Unlabeled Data: Light curves with 20 photometric measurements.

Two Options for Training Data

1. High SN: Many photometric measurements / light curve

- Pros: Accurately estimate features (eg period estimates correct)
- Cons: Training "looks different" than unlabeled data.

2. Training resembles Unlabeled: 20 photometric measurements

- Pros: Training "looks the same" as unlabeled.
- Cons: Features estimated incorrectly.

What Training Data To Use?

Unlabeled Data: Light curves with 20 photometric measurements.

Two Options for Training Data

1. High SN: Many photometric measurements / light curve

- Pros: Accurately estimate features (eg period estimates correct)
- Cons: Training "looks different" than unlabeled data.

2. Training resembles Unlabeled: 20 photometric measurements

- Pros: Training "looks the same" as unlabeled.
- Cons: Features estimated incorrectly.

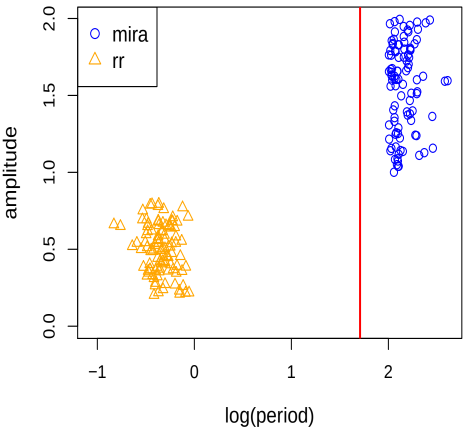
Training Data Should Resemble Unlabeled Data

Hypothetical Example:

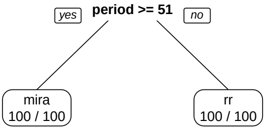
- Unlabeled Data: RR Lyrae and Miras with 20 **photometric measurements**
- Features: period and amplitude.
- Training 1: Light curves with >100 photometric measurements
- Training 2: Light curves with 20 photometric measurements

Classifier built on Training 1 Data

<u>Feature Distribution</u>



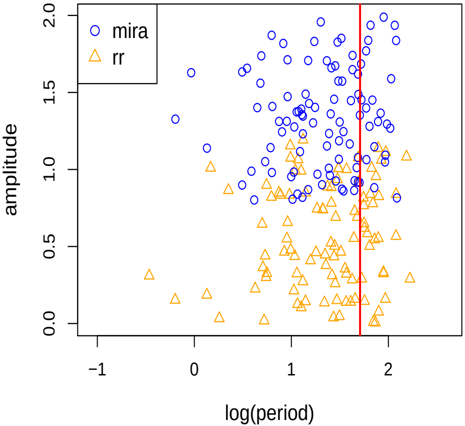
<u>CART Tree</u>



Conclusion: Seemingly Perfect Classification

Apply Classifier to Unlabeled Data

Feature Distribution



Confusion Matrix

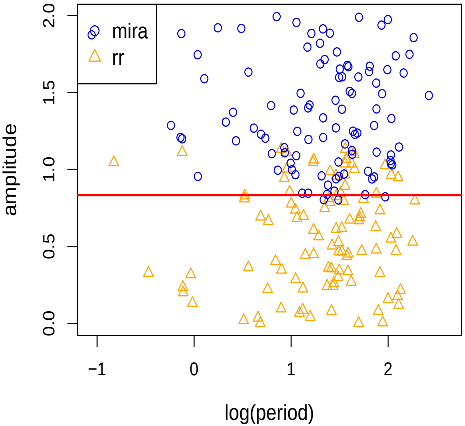
	Predicted	
Truth	mira	rr
mira	22	78
rr	28	72

Observations

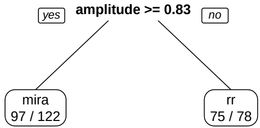
- classifier constructed using Training Data 1 used period to separate classes
- for poorly sampled unlabeled data, period does not separate classes (cannot compute period accurately)
- but amplitude is still useful for separating classes

Classifier built on Training 2 Data

Feature Distribution

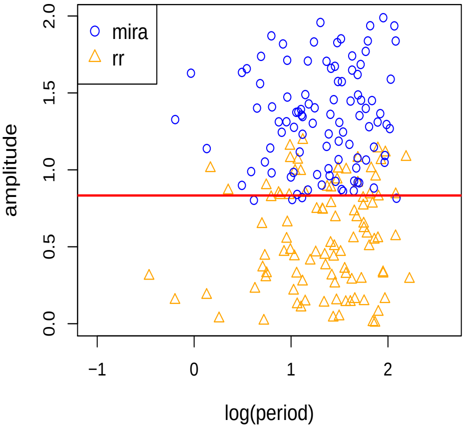


CART Tree



Apply Train 2 Classifier to Unlabeled Data

Feature Distribution



Confusion Matrix

	Predicted	
Truth	mira	rr
mira	96	4
rr	29	71

Conclusion: Much better performance.

Summary of Training Data Selection

- classifiers constructed on high SN data find class boundaries in high SN feature space
- these boundaries may not exist for low SN unlabeled data.
- downsampling high SN data to match unlabeled data SN can improve classifier performance
  - example of **domain adaptation** / **transfer learning**
  - Long et al. [4] for extensive discussion, methodology

Outline

Methodology: Statistical Classifiers

Methodology: CART Example with OGLE Data

Challenge 1: Selection of Training Data

Challenge 2: Classification versus Clustering

Conclusions and Opportunities

Recall Classification Example

Data:

- $\approx 100,000$ variable sources (large J–stetson) in M33
- ≈ 30 observations / source in I–band
- mix of Miras (O–rich/C–rich), SRVs, Cepheids, non–periodic sources, junk, etc.

Goals:

- find O–rich and C–rich Miras
- determine period–luminosity relationships for the Miras

Building Classifier for M33 is Difficult

OGLE Training Data

- downsample to match M33 cadence / photometric error
- select OGLE classes which match classes in M33

Evaluating Classifier Performance

- straightforward to measure error rate on training data
- how do we measure error rate on test?
- classification is only an intermediate step towards larger astronomy goals (specifically modeling of the light curve populations)

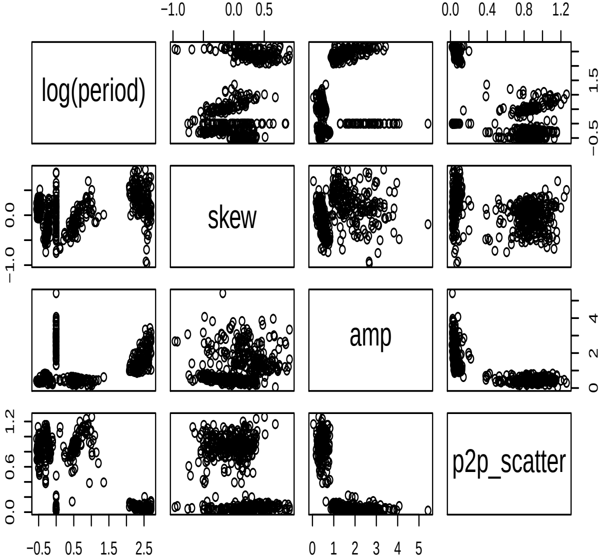
A Different Approach: Clustering

clustering: find groups (ie clusters) of objects in feature space

- compute feature distance between all objects
- find clusters where:
  - distance between objects within cluster is small
  - distance between objects in different clusters is large

clustering is different than classification: No training data. No objective measure of success.

OGLE Data



log(period)
skew
amp
p2p_scatter
−1.0
0.0
0.5
0.0
0.4
0.8
1.2
1.5
−0.5
0.0
−1.0
4
2
0
1.2
0.6
0.0
−0.5
0.5
1.5
2.5
0
1
2
3
4
5

Hierarchical Agglomerative Clustering Idea

Main Idea:

- every observation starts as own cluster
- iteratively merge "close" clusters together
- iterate until one giant cluster left

Method is

- **Hierarchical:** Each iteration produces a clustering, so do not specify number of clusters in advance.
- **Agglomerative:** Initially every observation in own cluster.

Hierarchical Agglomerative Clustering Pseudocode

- $N \leftarrow \{1, \ldots, n\}$
- $d_{ij} \leftarrow d(x_i, x_j) \; \forall \; i, j \in N$
- $C_{in} \leftarrow \{x_i\} \;\; \forall i \in N$
- **for** $k = n, \ldots, 2$**:**
  - $i, j \leftarrow \underset{\{i,j: i<j, \; i,j \in N\}}{\operatorname{argmin}} \; d_C(C_{ik}, C_{jk})$
  - $C_{i(k-1)} \leftarrow C_{ik} \cup C_{jk}$
  - $C_{l(k-1)} \leftarrow C_{lk} \; \forall l \neq i, j$ and $l \in N$
  - $N \leftarrow N \backslash \{j\}$

The $C_{\cdot k}$ are the k clusters in the k^{th} level of the hierarchy.

How to Merge Clusters (What is d_C?)

- Average Linkage

$$d_C(C_i, C_j) = \frac{1}{\#(C_i)\#(C_j)} \sum_{x \in C_i} \sum_{x' \in C_j} d(x, x')$$

- Complete Linkage

$$d_C(C_i, C_j) = \max_{x \in C_i, x' \in C_j} d(x, x')$$

- Single Linkage

$$d_C(C_i, C_j) = \min_{x \in C_i, x' \in C_j} d(x, x')$$

Constructing a Dendrogram

- At iteration k

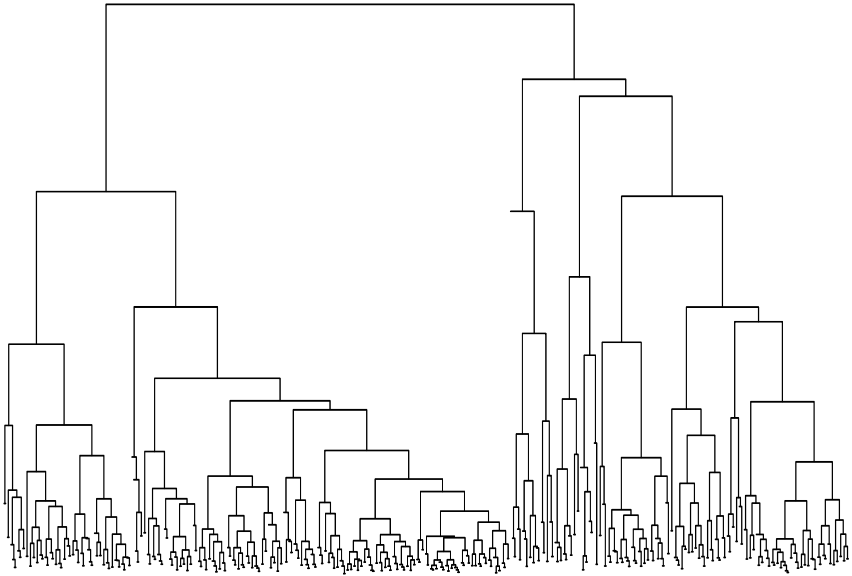
$$i, j \leftarrow \underset{\{i,j: i<j, i,j \in N\}}{\operatorname{argmin}} d_C(C_{ik}, C_{jk}).$$

- The "height" of this cluster merger is

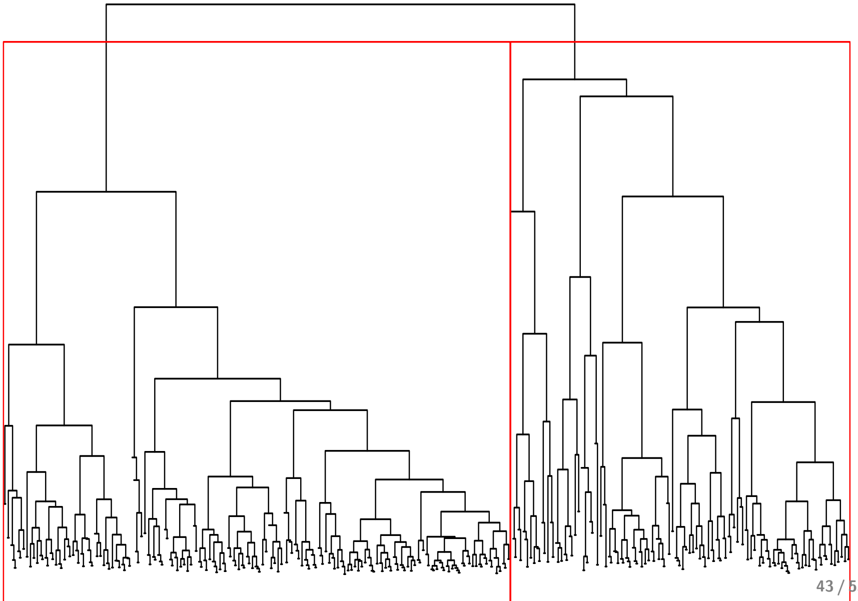
$$h_k = d_C(C_{ik}, C_{jk})$$

- The sequence $h_n, \ldots, h_2$ is monotonically increasing.
- Plot with heights of cluster mergers is a **dendrogram**.

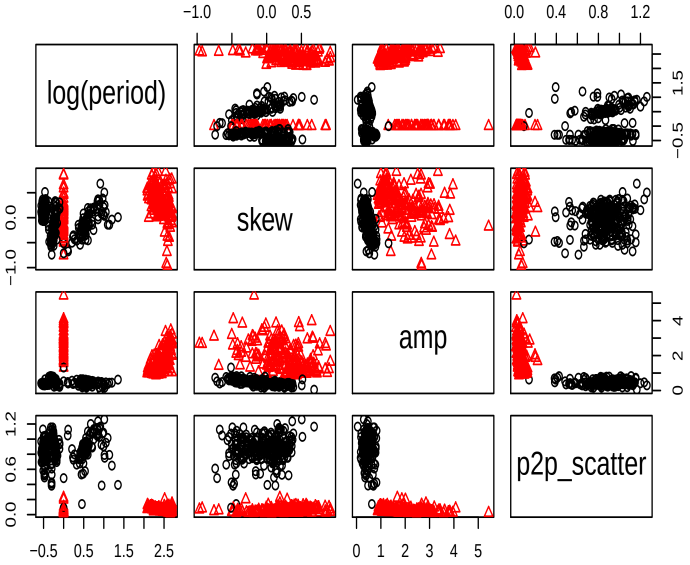
Clustering Dendrogram of OGLE Data



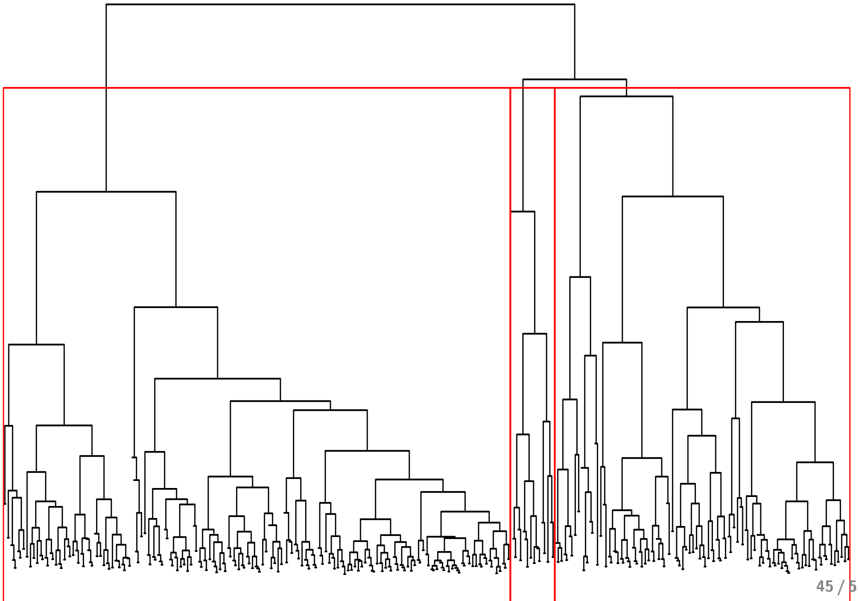
Two Clusters



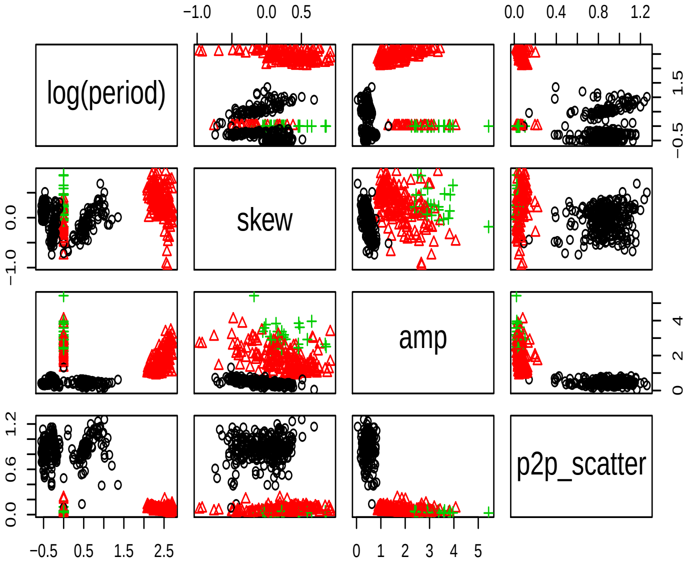
Two Clusters



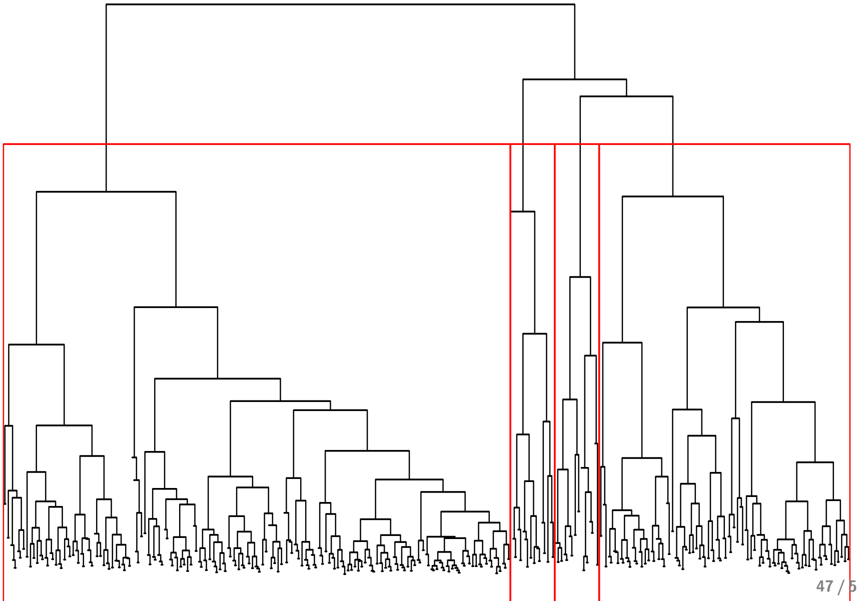
Three Clusters



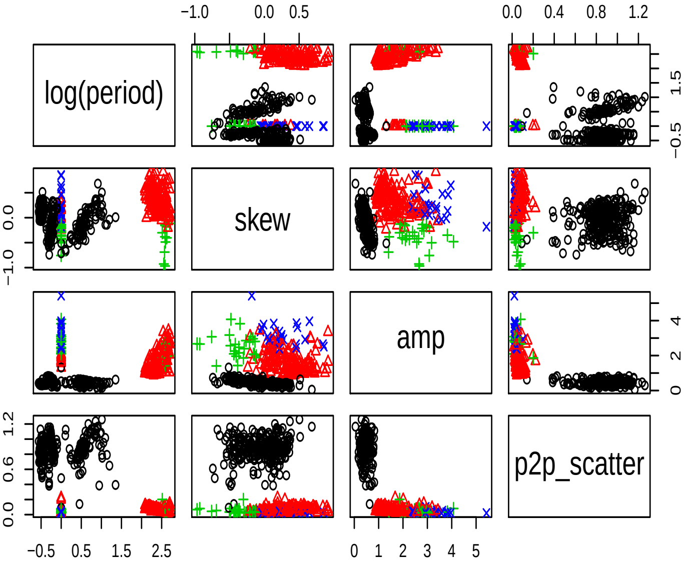
Three Clusters



Four Clusters



Four Clusters



Outlier Identification

- **“Supervised detection of anomalous light curves in massive astronomical catalogs” Nun, Pichara, Protopapas, Kim ApJ 2014 [6]**
  - Model voting distribution of random forest using Bayesian network. Outliers have unusual voting patterns.
- **“Discovery of Bright Galactic R Coronae Borealis and DY Persei Variables: Rare Gems Mined from ACVS” Miller, Richards, Bloom, et al. [5]**
  - Find rare R Cor Bor stars using random forest classifier, human analysis of light curves, and spectroscopic follow–up

More on Clustering and Outlier Detection

- many "knobs" in clustering methods
  - features
  - distance metric between features
  - hierarchical clustering, k–means, model based, etc.
- hard to statistically quantify successful clustering
  - may explain popularity of classification
- opinion: variable star "classification" is between the statistical concepts of clustering, classification, and outlier detection

Outline

Conclusions

- statistical classification
  1. select training data
  2. extract features
  3. build classifier
  4. apply classifier to unlabeled data
- training data should "look like" unlabeled data
- classification as practiced in statistics does not always fit perfectly with what astronomers want to do

Opportunities for Learning More / Project Ideas

- join Working Group 2 (WG2)
- compete in WG2 classification challenge (starting October?)
- machine learning tutorial on SDSS:
  `https://github.com/juramaga/clustering/blob/master/machine-learning-on-SDSS.ipynb`
- "Modeling Light Curves for Improved Classification" Faraway, J. Mahabal, A. et al. 2014 [3]
  Data available: `http://people.bath.ac.uk/jjf23/modlc/`
- outlier hunting in data set, eg OGLE
  `http://ogledb.astrouw.edu.pl/~ogle/CVS/`

Upcoming Topics in Time Domain

- astronomical motivation for time domain / variable sources
  - distance determination
  - period–luminosity relation
  - expansion of the universe
- feature extraction / modeling light curves
- example problem: mapping the Milky Way halo with RR Lyrae

Bibliography I

[1] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen.
Classification and regression trees.
Chapman & Hall/CRC, 1984.

[2] P. Dubath, L. Rimoldini, M. Süveges, J. Blomme, M. López, LM Sarro, J. De Ridder, J. Cuypers, L. Guy, I. Lecoeur, et al.
Random forest automated supervised classification of hipparcos periodic variable stars.
Monthly Notices of the Royal Astronomical Society, 414(3):2602–2617, 2011.

[3] Julian Faraway, Ashish Mahabal, Jiayang Sun, Xiaofeng Wang, Lingsong Zhang, et al.
Modeling light curves for improved classification.
arXiv preprint arXiv:1401.3211, 2014.

[4] J.P. Long, N. El Karoui, J.A. Rice, J.W. Richards, and J.S. Bloom.
Optimizing automated classification of variable stars in new synoptic surveys.
Publications of the Astronomical Society of the Pacific, 124(913):280–295, 2012.

[5] AA Miller, JW Richards, JS Bloom, SB Cenko, JM Silverman, DL Starr, and KG Stassun.
Discovery of bright galactic r coronae borealis and dy persei variables: Rare gems mined from acvs.
The Astrophysical Journal, 755(2):98, 2012.

[6] Isadora Nun, Karim Pichara, Pavlos Protopapas, and Dae-Won Kim.
Supervised detection of anomalous light curves in massive astronomical catalogs.
The Astrophysical Journal, 793(1):23, 2014.

[7] J.W. Richards, D.L. Starr, N.R. Butler, J.S. Bloom, J.M. Brewer, A. Crellin-Quick, J. Higgins, R. Kennedy, and M. Rischard.
On machine-learned classification of variable stars with sparse and noisy time-series data.
The Astrophysical Journal, 733:10, 2011.