Clustering Astronomical Data

5 Clustering Methods Compared

Dane Skinner Nick Hockensmith Kevin Park

Oregon State University

April 24, 2015

Outline

- Brief Introduction
- Questions of Interest
- Clustering Methods
- Conclusion and Final Thoughts

Introduction

- We have about 1500 training observations and 50,000 test observations of variable star data.
- The task is to cluster the data accurately.
- We focus on the training data because it allows us to compare the known clusters to the clusters the algorithms assign to the observations.

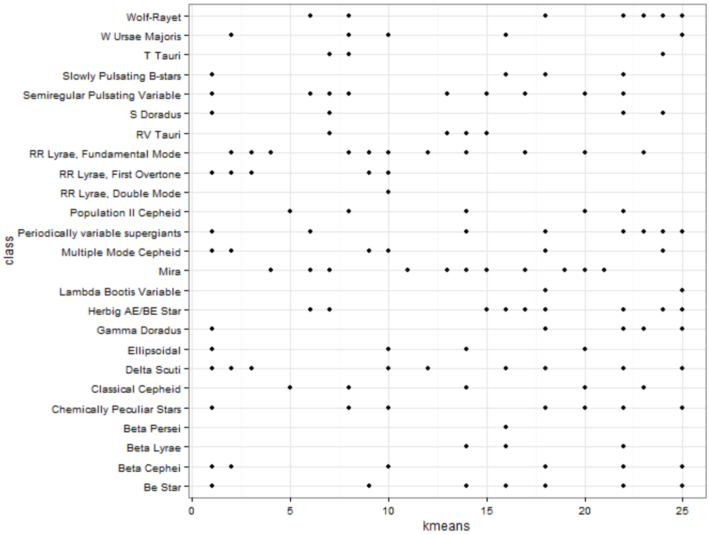
Questions of Interest

- What clustering algorithms perform the best?
- Does subsetting the data help (i.e. Does clustering in stages improve performance)?

K-Means Clustering

- Advantages:
  - Easy to implement.
  - Built in function in R is relatively quick.
- Disadvantages:
  - Poor performance when data has overlapping variable values.
  - NP-Hard problem.

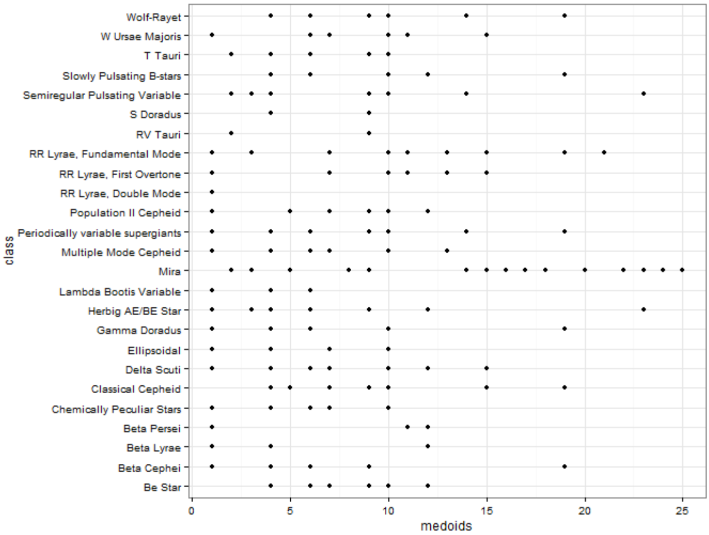
K-Means Clustering Results



K-Medoids Clustering

- Advantages:
  - Improved performance over K-means.
  - Freedom to choose distance metric (typically any ℓ_p metric) although R only allows ℓ_1 or ℓ_2.
- Disadvantages:
  - Performance still suffers when data has overlapping variable values.
  - Another NP-Hard problem.

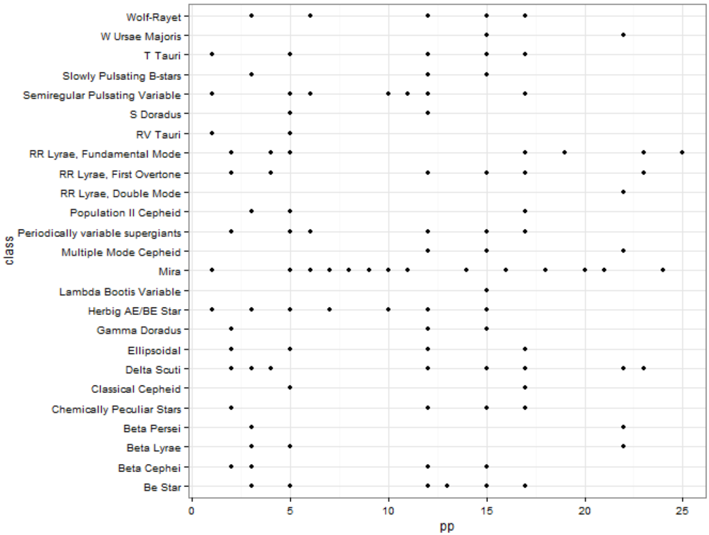
K-Medoids Clustering Results



K-Means ++

- Advantages:
  - Iterative process yields improved performance over K-medoids.
  - Algorithm checks multiple possible cluster centers and returns best cluster assignments.
- Disadvantages:
  - Iterative process takes time.
  - Algorithm may miss ideal solution since there are so many possible starting configurations.

K-Means++ Clustering Results



Spectral Clustering Methods

- Chose the complete linkage over the single or average linkages
- All 85 explanatory variables were used for hierarchical clustering
- Each column variable was normalized to the interval $[0, 1]$ via

$$x_i^* = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

- Advantages:
  - Adjusts column variables to a notionally common scale
  - "Better" spread in observation assignment to different clusters
  - Quick run time
- Disadvantages
  - Algorithm still wants to push many observations into a small number of clusters
  - Artificial scaling of units may or may not be a good thing

Tables

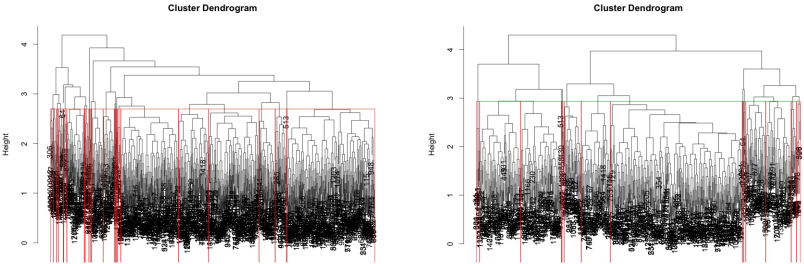
C. ID	1	2	3	4	5	6	7	8	9	10	11	12
Obs.	196	38	11	171	298	102	39	9	38	39	55	21
C. ID	13	14	15	16	17	18	19	20	21	22	23	24
Obs.	1	13	11	17	3	6	11	4	4	1	3	6

Table : Farthest-Neighbor on Non-subsetted Training Data: Normalized (3)

C. ID	1	2	3	4	5	6	7	8	9	10	11	12
Obs.	856	22	7	4	37	81	4	11	45	1	2	1
C. ID	13	14	15	16	17	18	19	20	21	22	23	24
Obs.	6	2	12	7	2	1	1	1	1	1	1	3

Table : Farthest-Neighbor on Non-subsetted Training Data: Standardized (1)

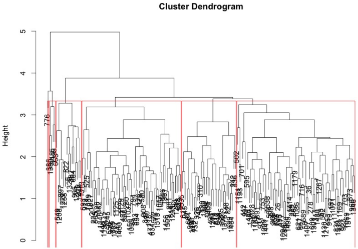
Dendrograms



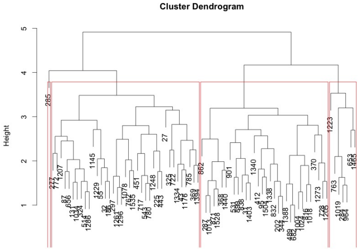
(a) All Star types

(b) Pulsating Star types only

Dendrograms

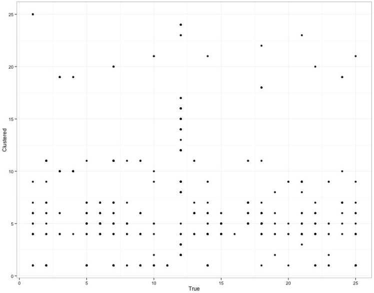


(c) Eruptive Star types only

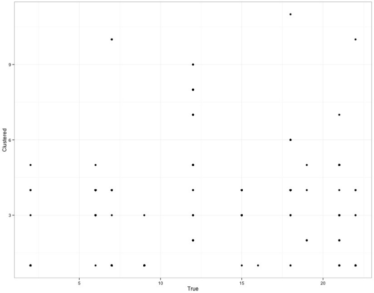


(d) Multi-Star types only

Spectral Clustering Results

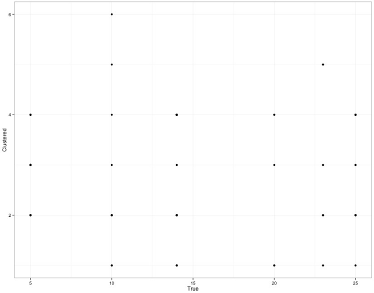


(e) All Star types

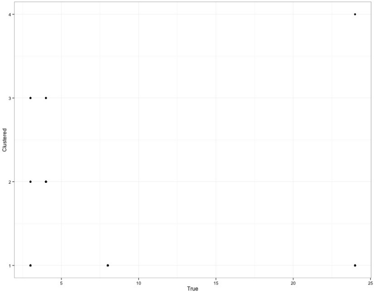


(f) Pulsating Star types only

Spectral Clustering Results



(g) Eruptive Star types only



(h) Multi-Star types only

Conclusions and Final Thoughts