Classification of Transients Using Machine Learning Methods

Linn Abraham & Robin Jacob Roy

August 10, 2023

Outline

Objective

The main objective of the project is to develop an automated classification technique to classify astronomical transients from an incoming stream of time series data using machine learning algorithms.

Some of the Variable star categories

- EW
- EA
- Blazkho
- RRab
- RRc
- RS CVn
- HADS
- ACEP
- Hump
- ELL

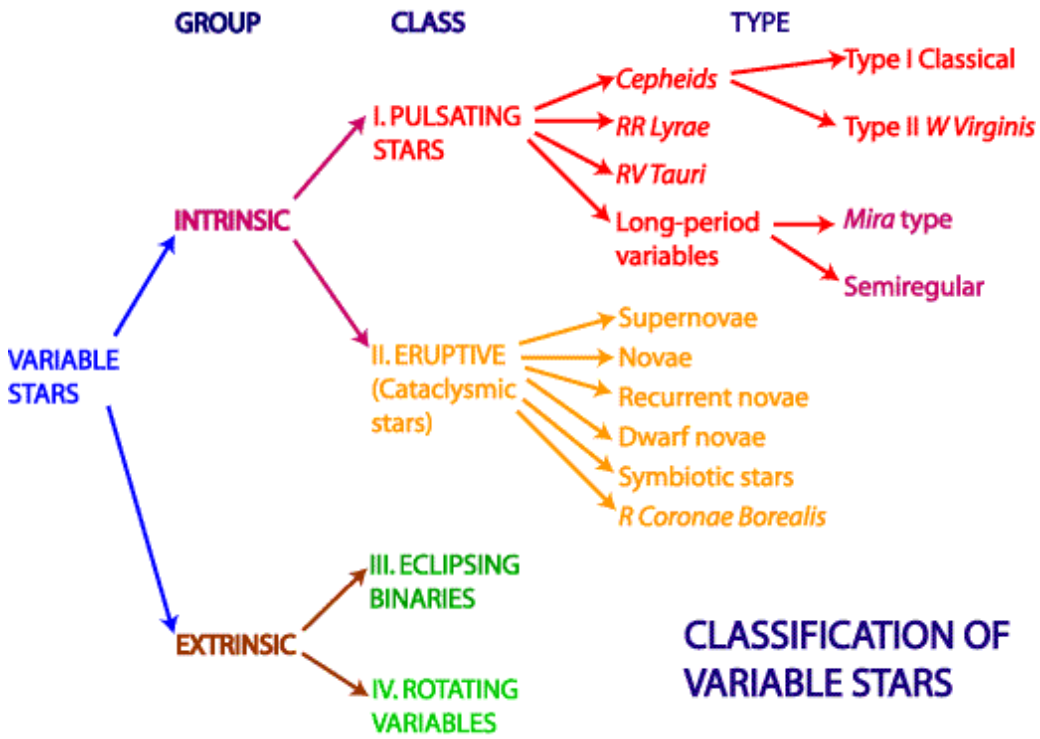


Figure: eclipsing binary system, including a Cepheid variable

Challenges

Some of the challenges of using traditional methods are:

- Big Data

Challenges

Some of the challenges of using traditional methods are:

- Big Data
- Real time decision making

Challenges

Some of the challenges of using traditional methods are:

- Big Data
- Real time decision making
- Require a greater understanding of the underlying astrophysics

Challenges

Some of the challenges of using traditional methods are:

- Big Data
- Real time decision making
- Require a greater understanding of the underlying astrophysics

Some of the challenges in the machine learning approach:

- Sparse data

Challenges

Some of the challenges of using traditional methods are:

- Big Data
- Real time decision making
- Require a greater understanding of the underlying astrophysics

Some of the challenges in the machine learning approach:

- Sparse data
- Noisy Data

Challenges

Some of the challenges of using traditional methods are:

- Big Data
- Real time decision making
- Require a greater understanding of the underlying astrophysics

Some of the challenges in the machine learning approach:

- Sparse data
- Noisy Data
- Identifying features

CRTS Data

- We have used publicly available data from the Catalina Real Time Transient Survey

CRTS Data

- We have used publicly available data from the Catalina Real Time Transient Survey
- CRTS data comes from the three telescopes run by the Catalina Sky Survey.

CRTS Data

- We have used publicly available data from the Catalina Real Time Transient Survey
- CRTS data comes from the three telescopes run by the Catalina Sky Survey.
  - The Mt. Lemmon Survey 1.5m Cassegrain reflector

CRTS Data

- We have used publicly available data from the Catalina Real Time Transient Survey
- CRTS data comes from the three telescopes run by the Catalina Sky Survey.
  - The Mt. Lemmon Survey 1.5m Cassegrain reflector
  - Catalina Sky Survey 0.7m Schmidt

CRTS Data

- We have used publicly available data from the Catalina Real Time Transient Survey
- CRTS data comes from the three telescopes run by the Catalina Sky Survey.
  - The Mt. Lemmon Survey 1.5m Cassegrain reflector
  - Catalina Sky Survey 0.7m Schmidt
  - Siding Springs Survey 0.5m Schmidt

CRTS Data

- We have used publicly available data from the Catalina Real Time Transient Survey
- CRTS data comes from the three telescopes run by the Catalina Sky Survey.
  - The Mt. Lemmon Survey 1.5m Cassegrain reflector
  - Catalina Sky Survey 0.7m Schmidt
  - Siding Springs Survey 0.5m Schmidt

CRTS Data

- We have used publicly available data from the Catalina Real Time Transient Survey
- CRTS data comes from the three telescopes run by the Catalina Sky Survey.
  - The Mt. Lemmon Survey 1.5m Cassegrain reflector
  - Catalina Sky Survey 0.7m Schmidt
  - Siding Springs Survey 0.5m Schmidt
- More than 800 million lines of data

CRTS Data

- We have used publicly available data from the Catalina Real Time Transient Survey
- CRTS data comes from the three telescopes run by the Catalina Sky Survey.
  - The Mt. Lemmon Survey 1.5m Cassegrain reflector
  - Catalina Sky Survey 0.7m Schmidt
  - Siding Springs Survey 0.5m Schmidt
- More than 800 million lines of data
- A total of 47,000 sources

The photometric data consists of:

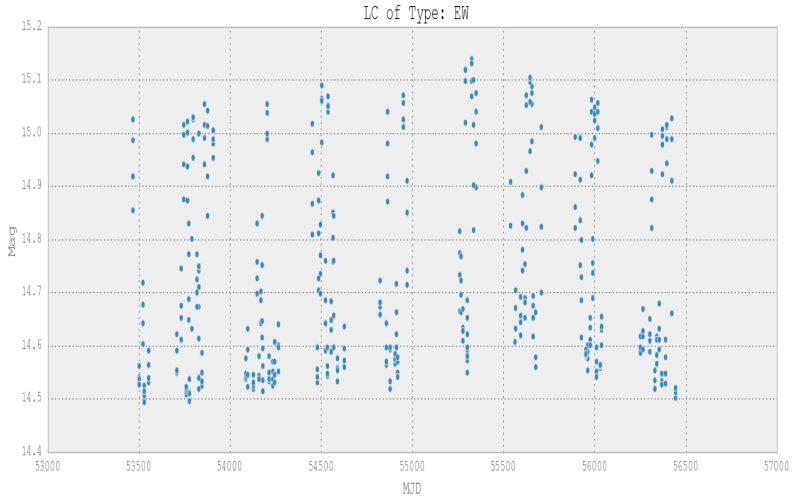
1. ID
2. Time
3. Magnitude information
4. Mag error
5. RA
6. Dec

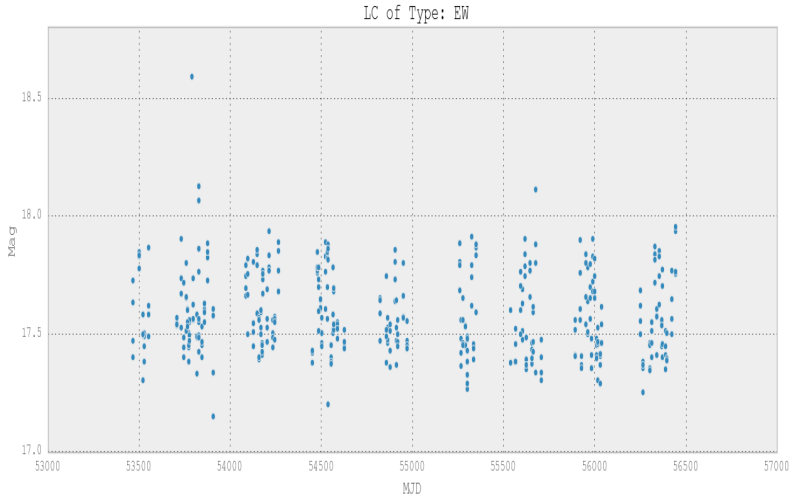
Python Packages

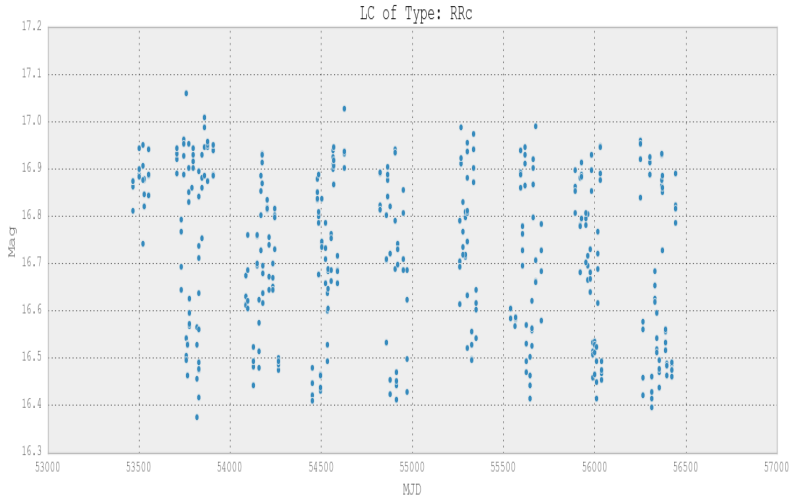
Here we have implemented machine learning algorithms as well as all the data processing in the python programming language.
The important libraries we have used for the purpose are:

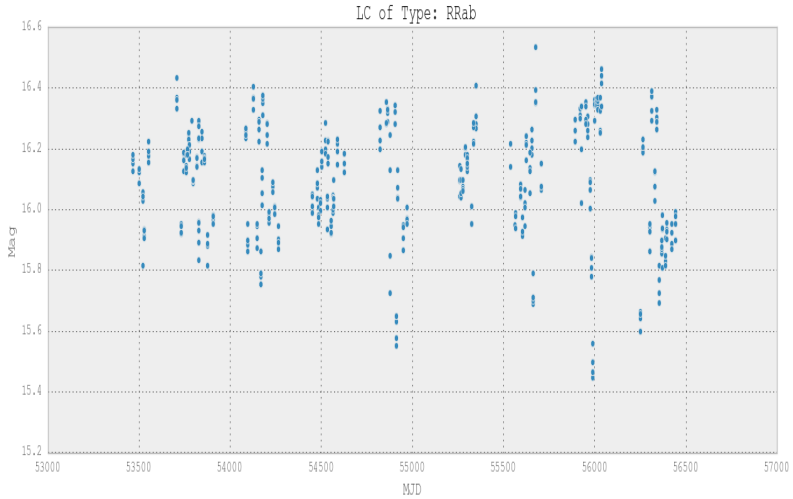
- Pandas - for creating and manipulating dataframes
- Scikit learn-machine learning
- Matplotlib- provides a matlab like functionality
- FATS- for feature extraction

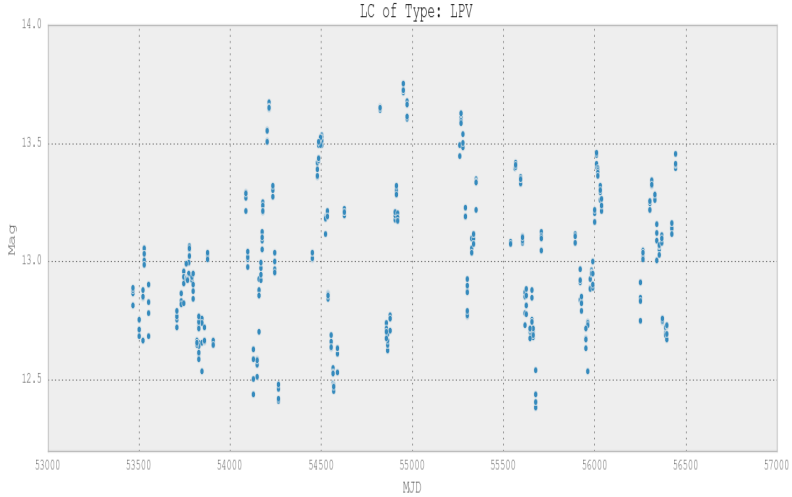
Light curves

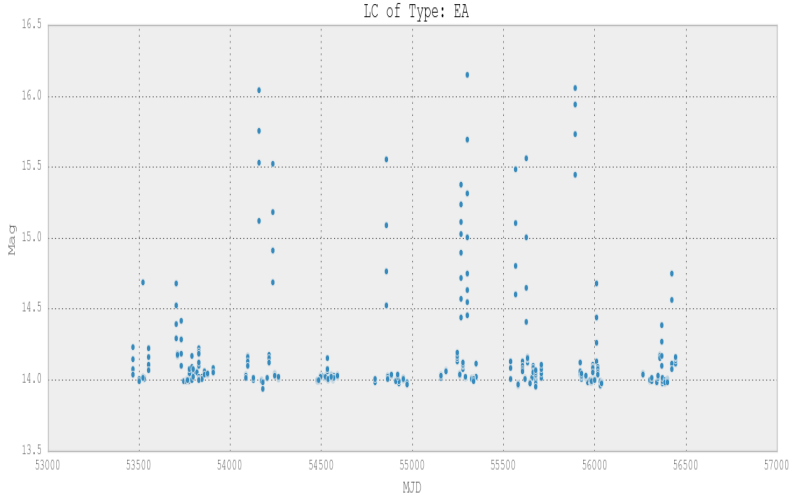


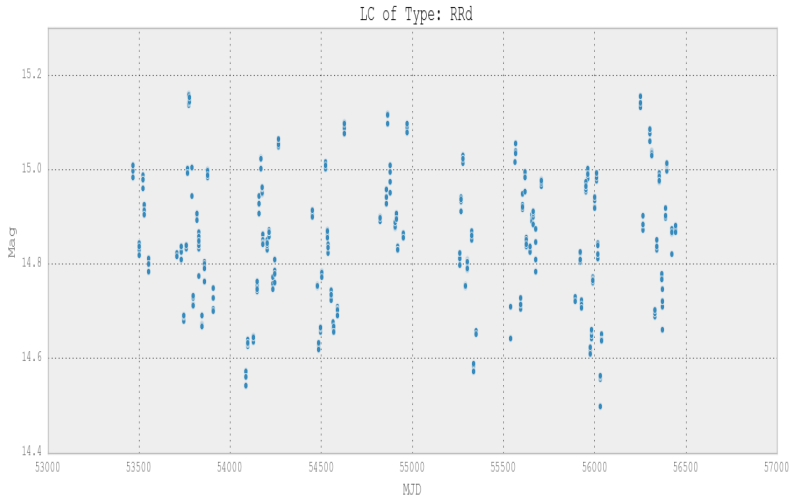


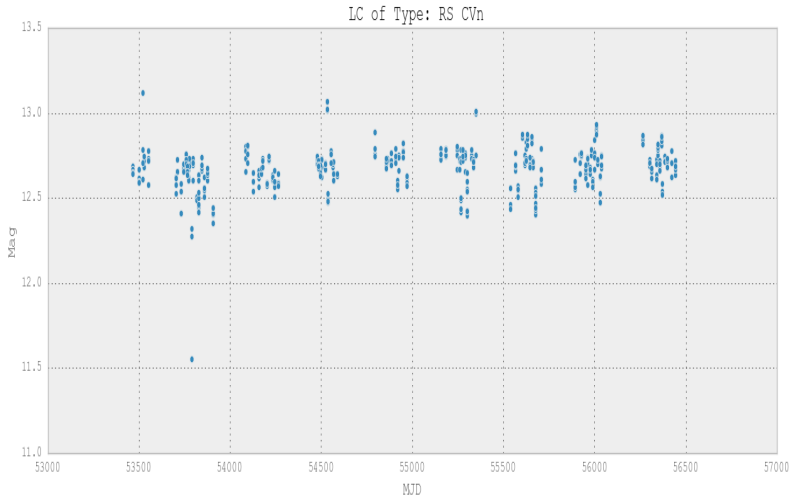





LC of Type: LPV
Mag
MJD


LC of Type: EA
Mag
MJD

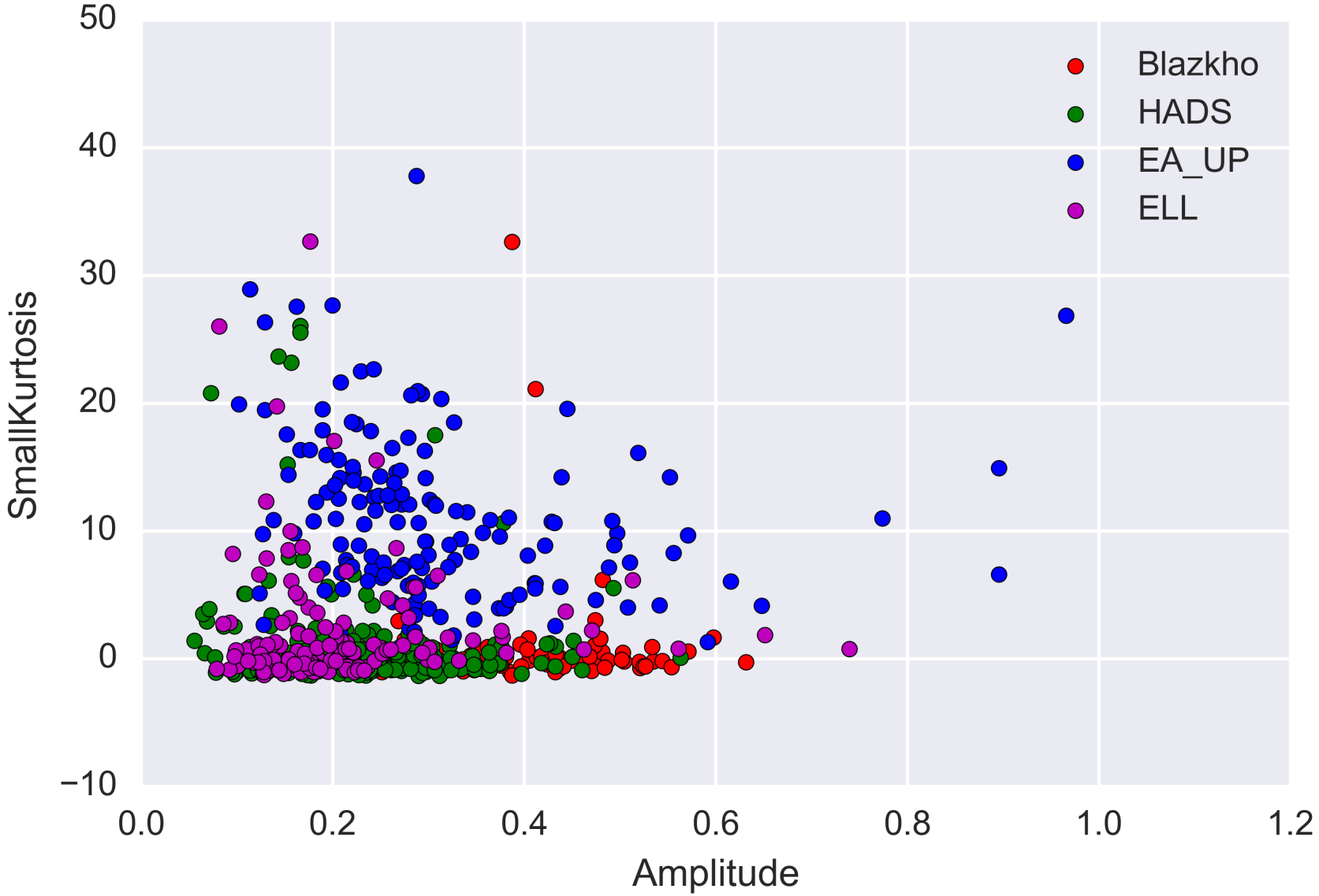

LC of Type: RRd
Mag
MJD


LC of Type: RS CVn
Mag
MJD

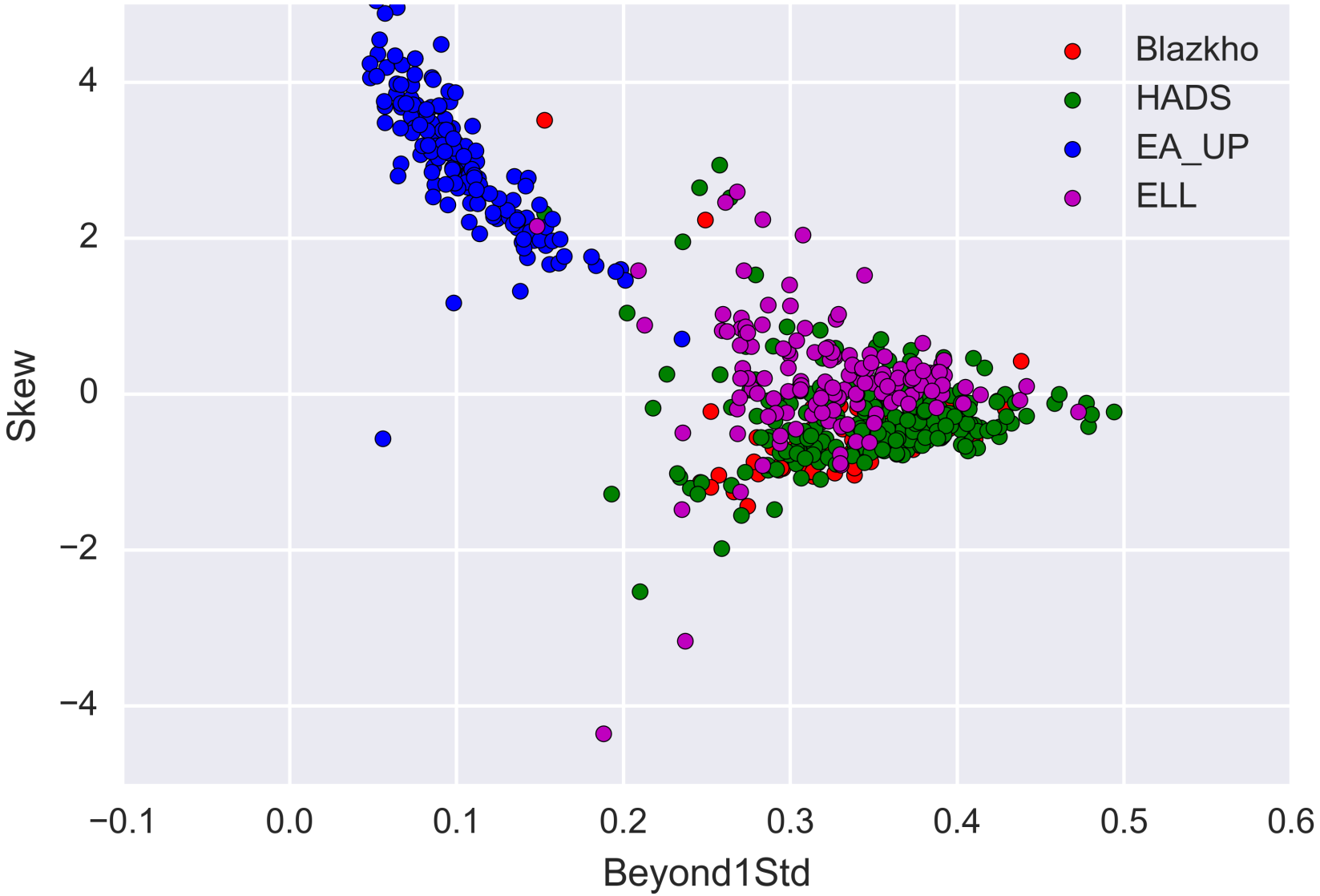
Features extracted from Light curves

1. Amplitude
2. Autocor_length
3. Beyond1Std
4. LinearTrend
5. MaxSlope
6. MedianAbsDev
7. MedianBRP
8. PercentAmplitude
9. Skew
10. SmallKurtosis
11. Std
12. StetsonK

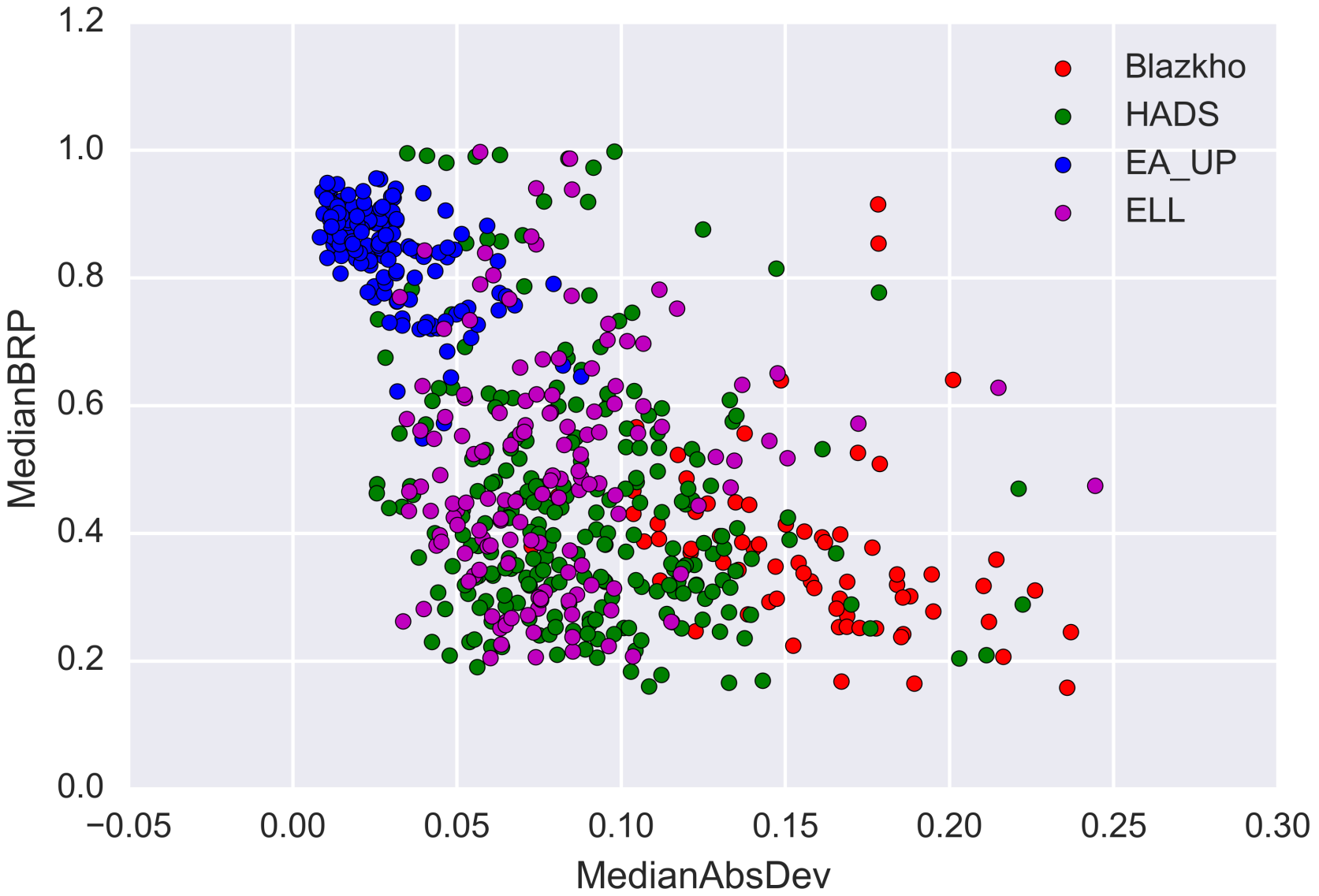
Feature plots



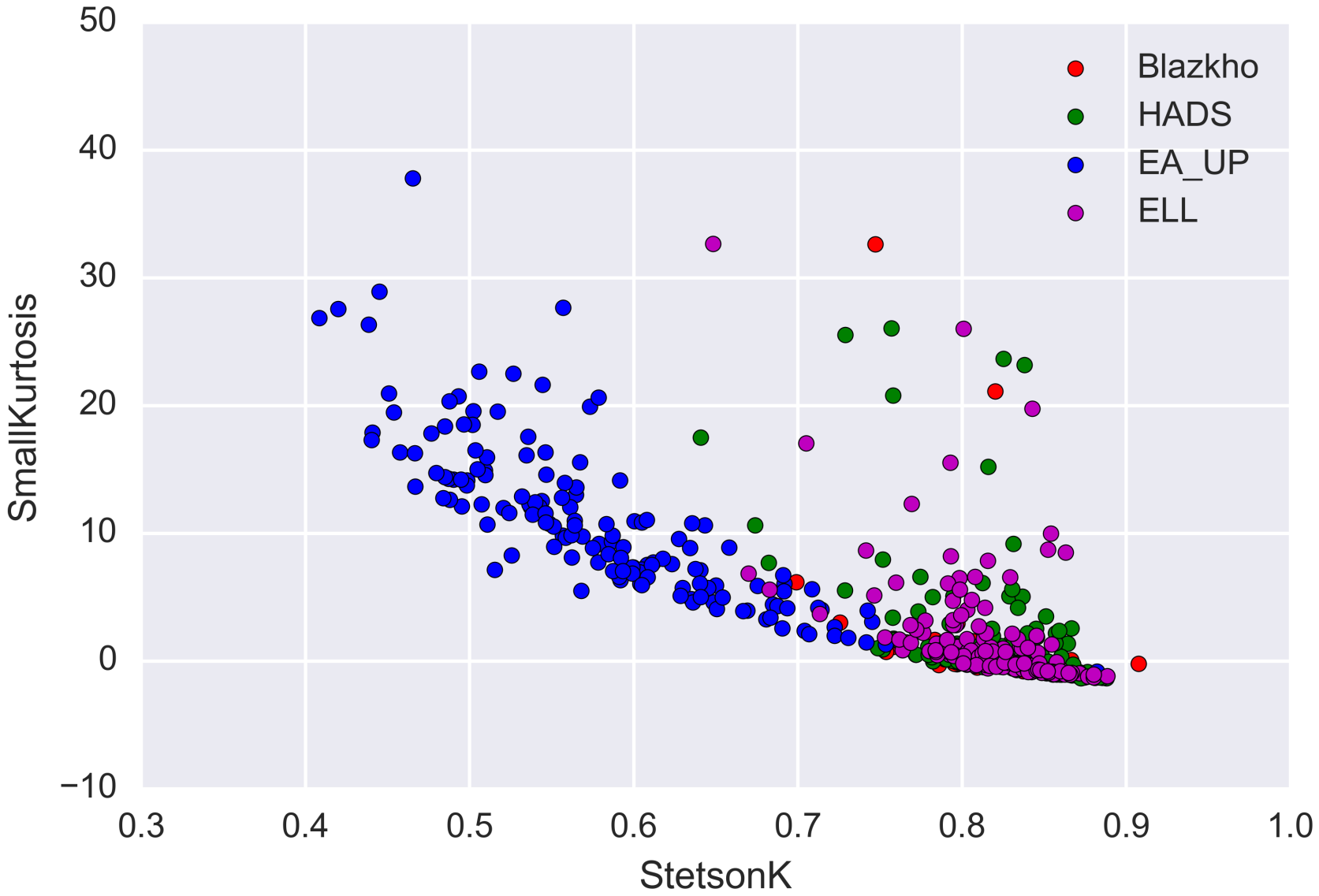
Feature plots



Feature plots



Feature plots

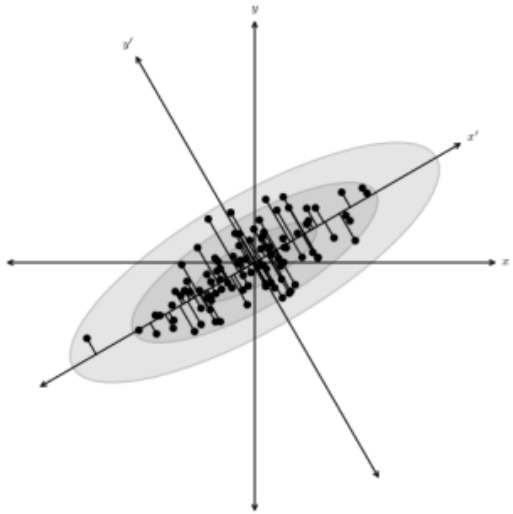


The main goals of PCA

- It aims to reduce the dimensionality of the data by detecting the correlation between the original features.
  - Thus it provides 'new features' which could be different combinations of the original ones.
- Dimensionality reduction also helps reduce over fitting of the ML algorithm to the training data.
- It helps the ML algorithm to run faster.

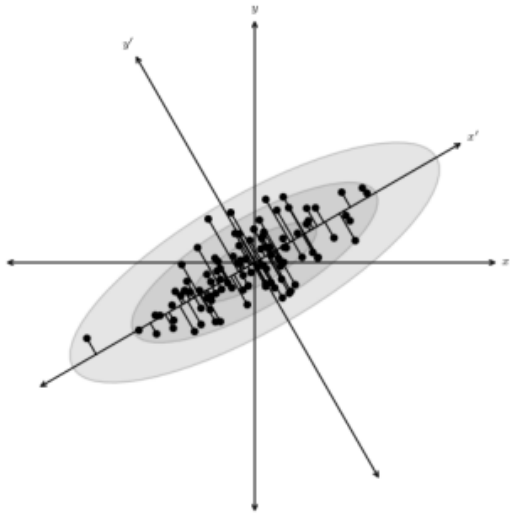
Dimensionality reduction using PCA

Finding the directions of maximum variance in high-dimensional data and projecting it onto a smaller dimensional subspace while retaining most of the information.



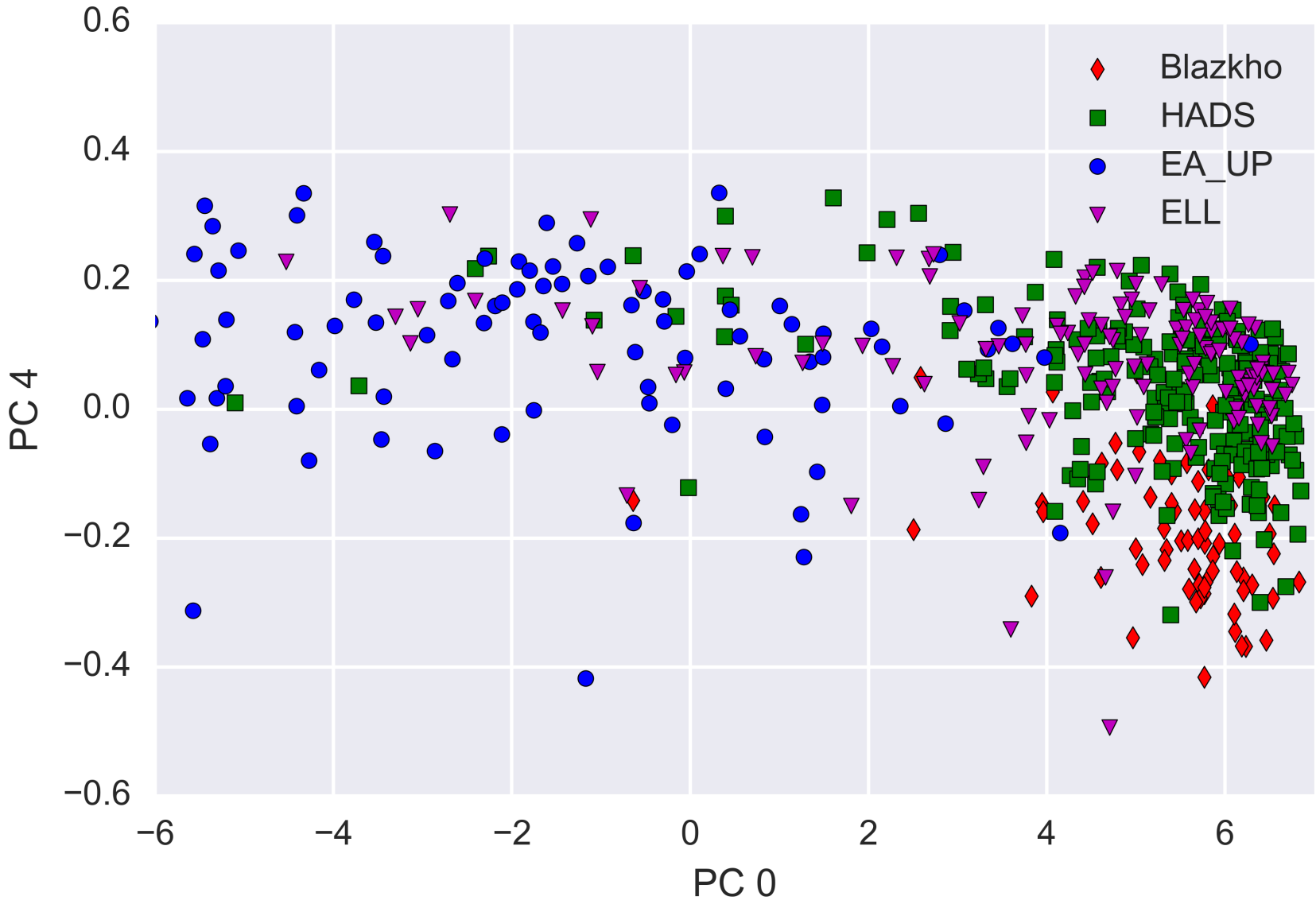
Dimensionality reduction using PCA

Finding the directions of maximum variance in high-dimensional data and projecting it onto a smaller dimensional subspace while retaining most of the information.

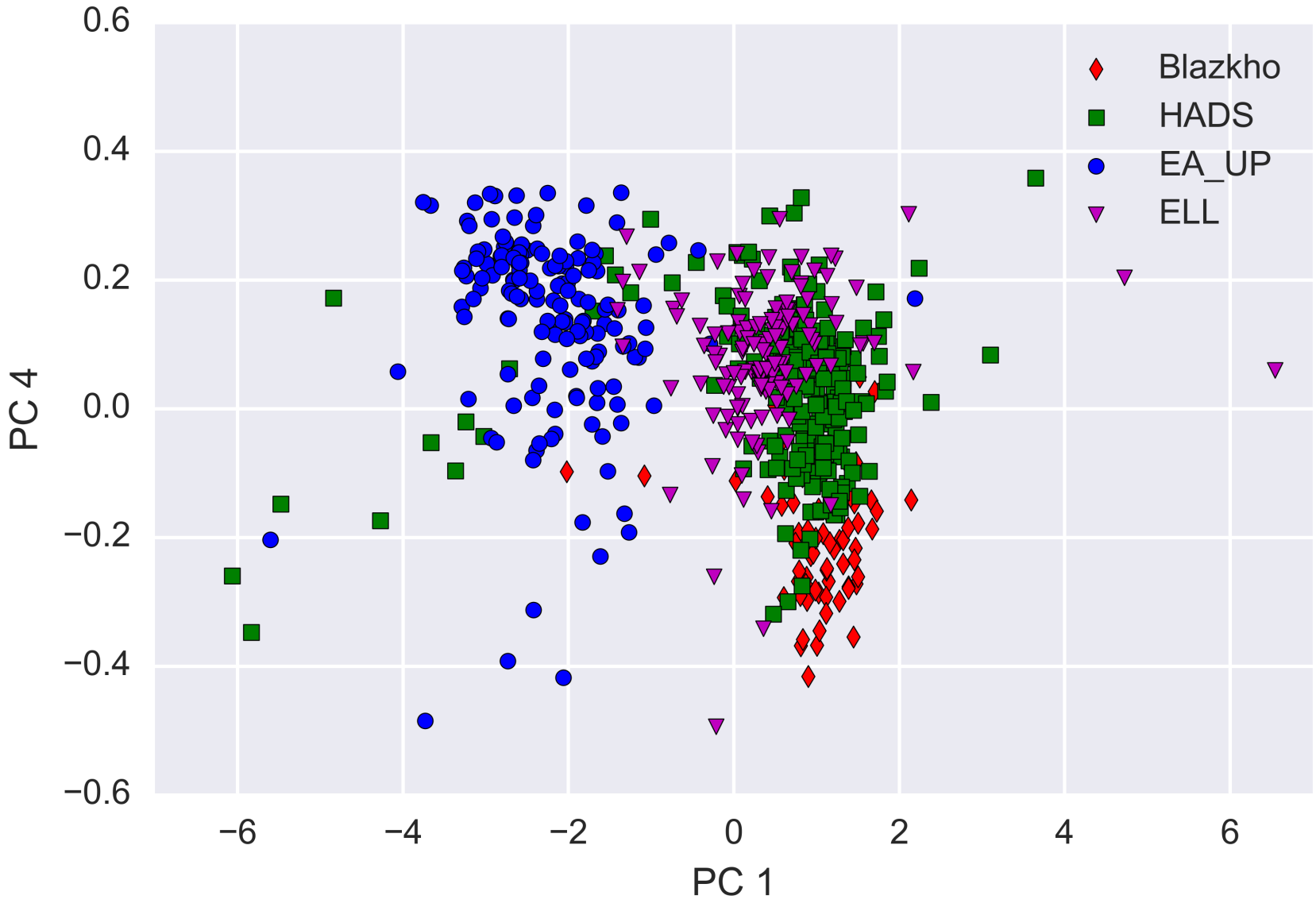


Principal Component Analysis Technique:
Consists of finding the eigenvalues λ_j and eigenvectors e_j of the data's correlation matrix $\Sigma = Y^T Y$ where $Y = X - \mu X$ and X is an $N \times M$ matrix of data points.μX is the empirical mean value of the data

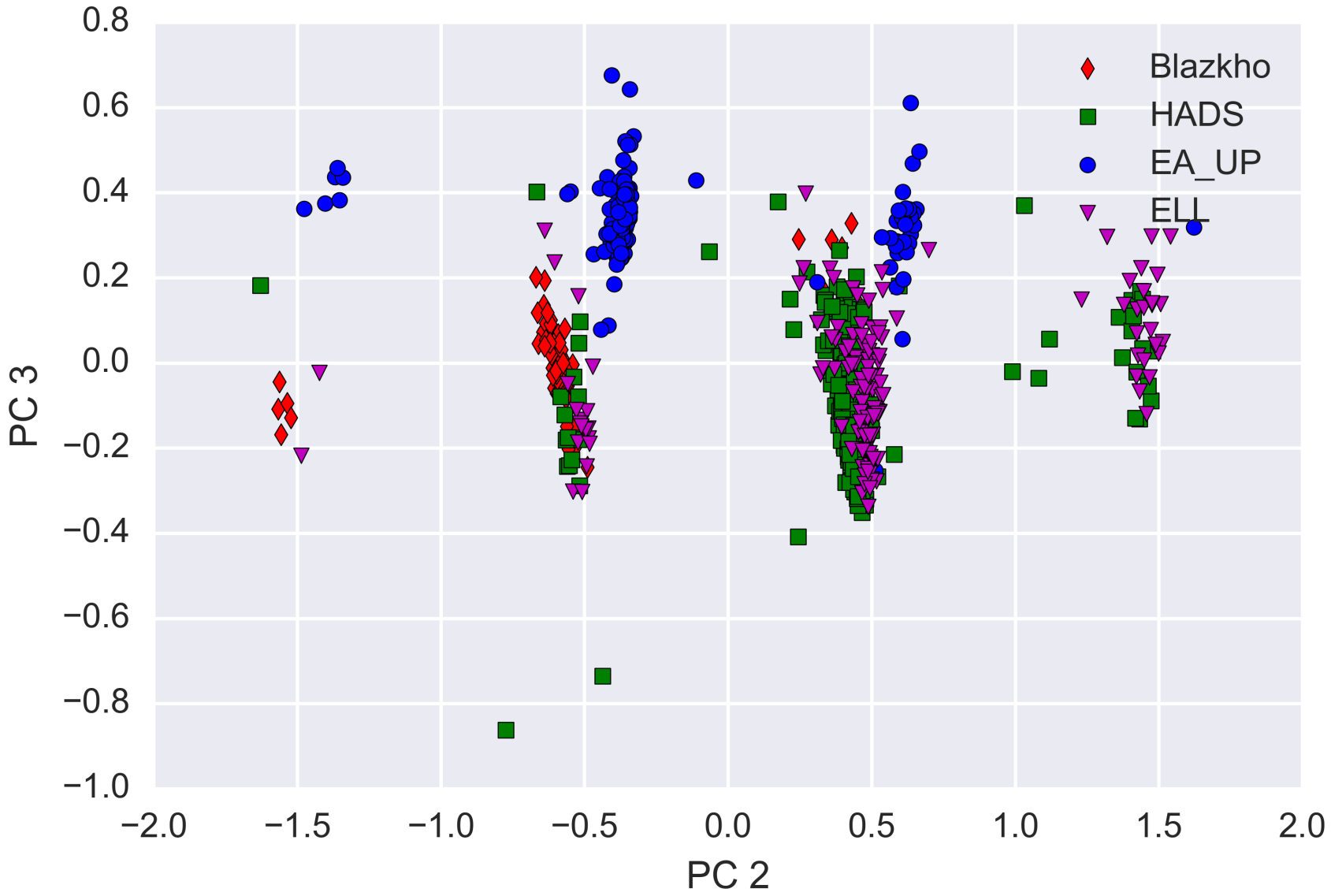
PCA plots



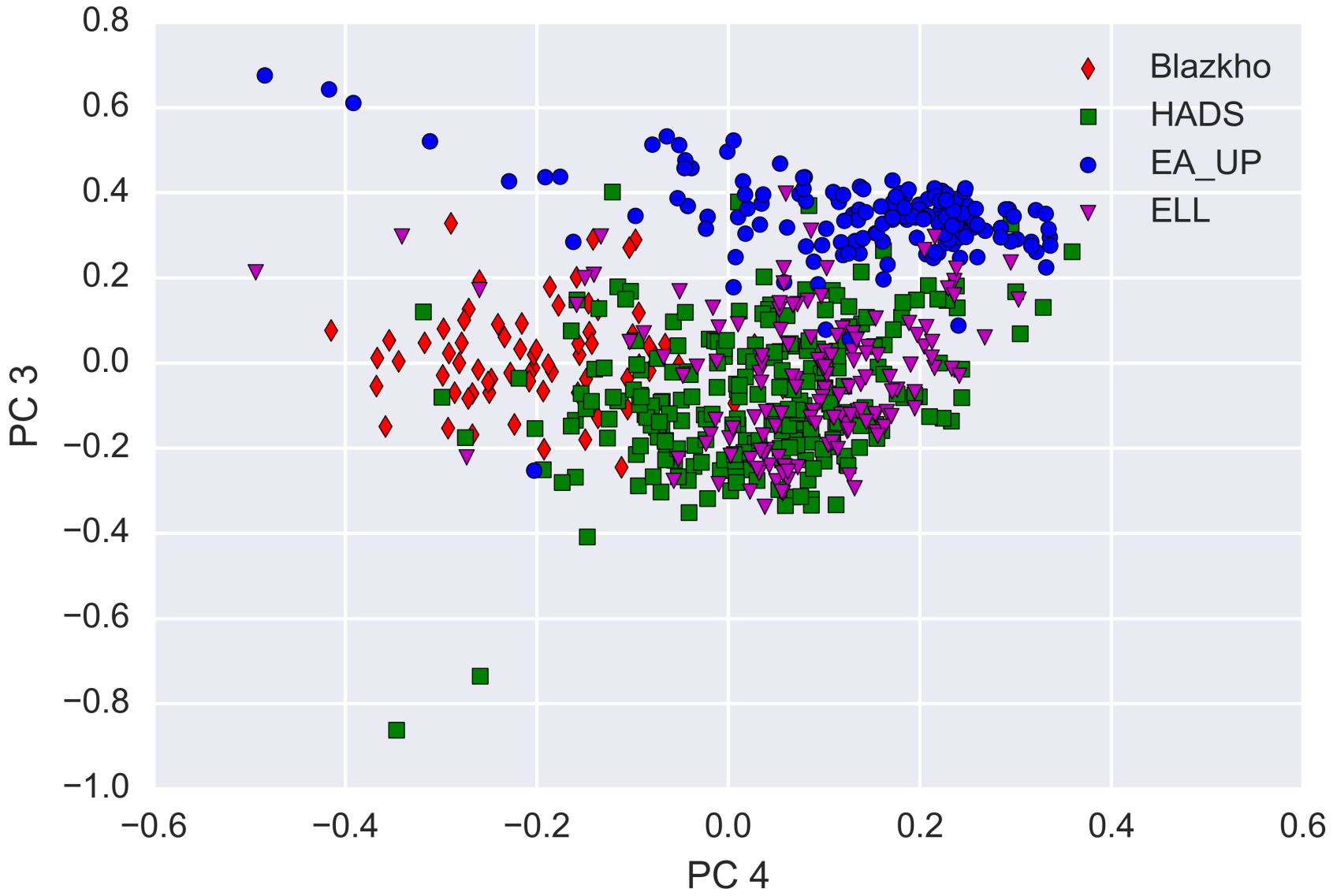
PCA plots



PCA plots


Blazkho
HADS
EA_UP
EU
PC 3
PC 2
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

PCA plots


Blazkho
HADS
EA_UP
ELL
PC 3
PC 4
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6

Machine Learning

Definition

Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959).

Machine Learning

Definition

Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959).

- Grew out of work in Artificial Intelligence

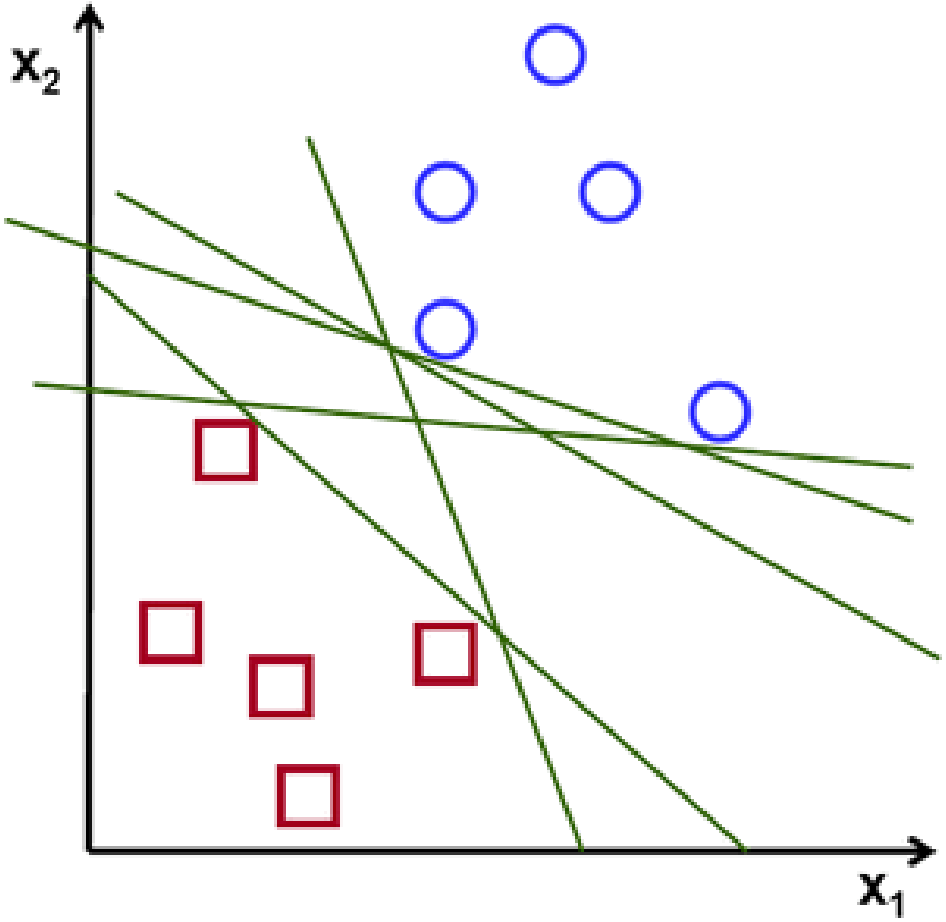
Machine Learning

Definition

Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959).

- Grew out of work in Artificial Intelligence
- Applications are diverse: Database mining, handwriting recognition, Natural Language Processing (NLP), Computer Vision.

Support Vector Machines



Support Vector Machines

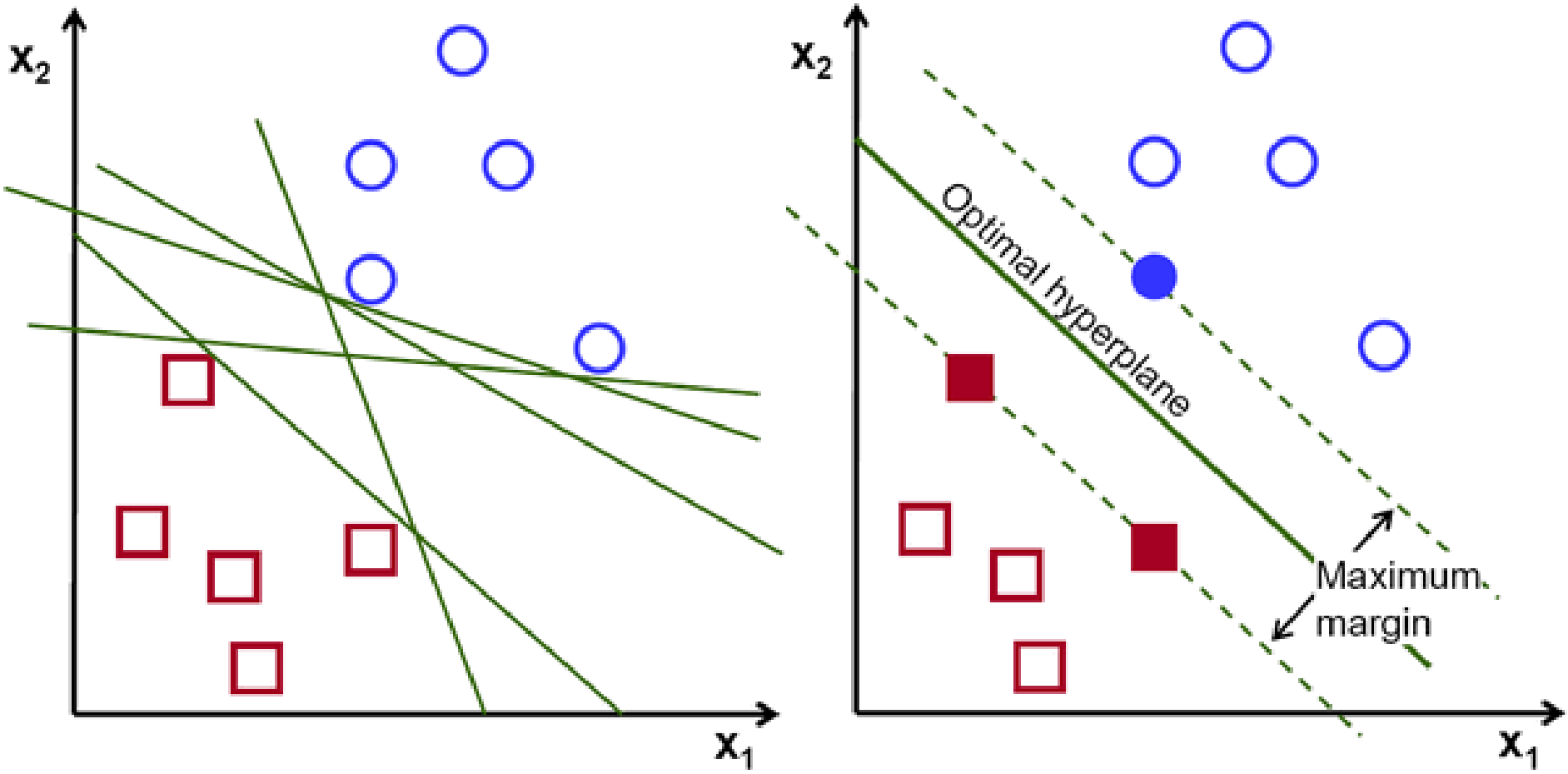


Figure: *Adapted from*
http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html

Results

Table: Group-wise classification accuracy

	Type	Precision	Recall	f1-score	Support
0	ACEP	0.42	0.73	0.53	22
1	Cep-II	0.65	0.64	0.65	53
2	EA_UP	0.97	1.00	0.98	63
3	ELL	0.65	0.84	0.73	49
4	PCEB	0.67	0.15	0.24	40
	avg / total	0.72	0.70	0.68	227
5	Blazkho	0.80	0.30	0.43	27
6	HADS	0.67	0.86	0.76	101
7	LPV	0.97	0.96	0.97	197
8	RRd	0.82	0.72	0.77	208
9	beta Lyrae	0.76	0.88	0.82	109
	avg / total	0.83	0.83	0.82	642
10	EA	0.87	0.92	0.89	1881
11	RRc	0.85	0.91	0.88	2133
12	RS CVn	0.79	0.47	0.59	628
	avg / total	0.85	0.85	0.84	4642
13	EW	0.95	1.00	0.97	12227
14	RRab	0.88	0.38	0.53	979
	avg / total	0.95	0.95	0.94	13206

Conclusion

1. Obtain the time series data for variables from CRTS
2. Extract features from the light curve data.
3. Study the effect of different features on the classification of variable types.
4. Use dimensionality reduction algorithms like PCA to select features
5. Apply supervised machine learning methods to classify the transients with an average accuracy of 83.75 %.

Further work to be done ...

- Read about the existing methods of classification from papers (of CRTS etc.)
- Read about hierarchical classification.
- Use multi dimensional plotting softwares like ggob to view the feature space.
- Find and remove outliers in the data. Use a suitable criterion to exclude points with large error margins.
- Study about the different features which are relevant to the case of variable stars.
- Use PCA to study the contribution of different features to the variance in the data and also to find the correlation between different features.
- Compare the effect of doing PCA transformation on the training data and that without.
- Study SVM in depth (using non linear kernels etc.).
- Compare the accuracy of other machine learning methods.