Amadeus Exoplanete Data Analysis

Jordi Nin, Marc Solé, Dino Ienco

nin,msole@ac.upc.edu
dino.ienco@teledetection.fr

LIRMM-CNRS and DAC-UPC

Data set description

- Number of instances: 97,718
- Number of attributes: 58 + id + position + label
- Possible labels: C, N, G, L, F, X and M $\rightarrow$ only C and N are valid status
- Number of classified instances: 4 C, 16 N $\rightarrow$ 0.02%
- There are some rows without a non-null flag attribute, concretely a 'space' ... control instances?

Attribute description

At least at the UPC, we have no idea about the meaning...

We have discarded the error measurement, the id, the space position and the attributes with most of their values as missing values

We still have 31 attributes:

mag_b, mag_v, mag_r, mag_i, mahalanobisHF, prob1HF, code1HF, pf1HF, f1HF, Amp11*100HF, Amp12*100HF, Amp13*100HF, Amp14*100HF, phdiff12HF, phdiff13HF, phdiff14HF, varredHF, mahalanobis, prob1, code1, f1, Amp11*100, Amp12*100, Amp13*100, Amp14*100, phdiff12, phdiff13, phdiff14, varred, period, epoch, duration, depth

The curse of dimensionality!

General problem

The curse of dimensionality refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces.

Machine learning

Problems that involve learning a "state-of-nature" (maybe an infinite distribution) from a finite number of data samples in a high-dimensional feature space with each feature having a number of possible values, **an enormous amount of training data are required** to ensure that there are several samples with each combination of values.

Some numbers...

Considering the curse of dimensionality problem and the number of attributes

$\downarrow$

We need at least 30 positive examples to split the complete space into two subspaces, one for the positive and another for the negative classes

Problem

We only have 4 positive instances ... then our space can have 4 dimensions at most!!!!

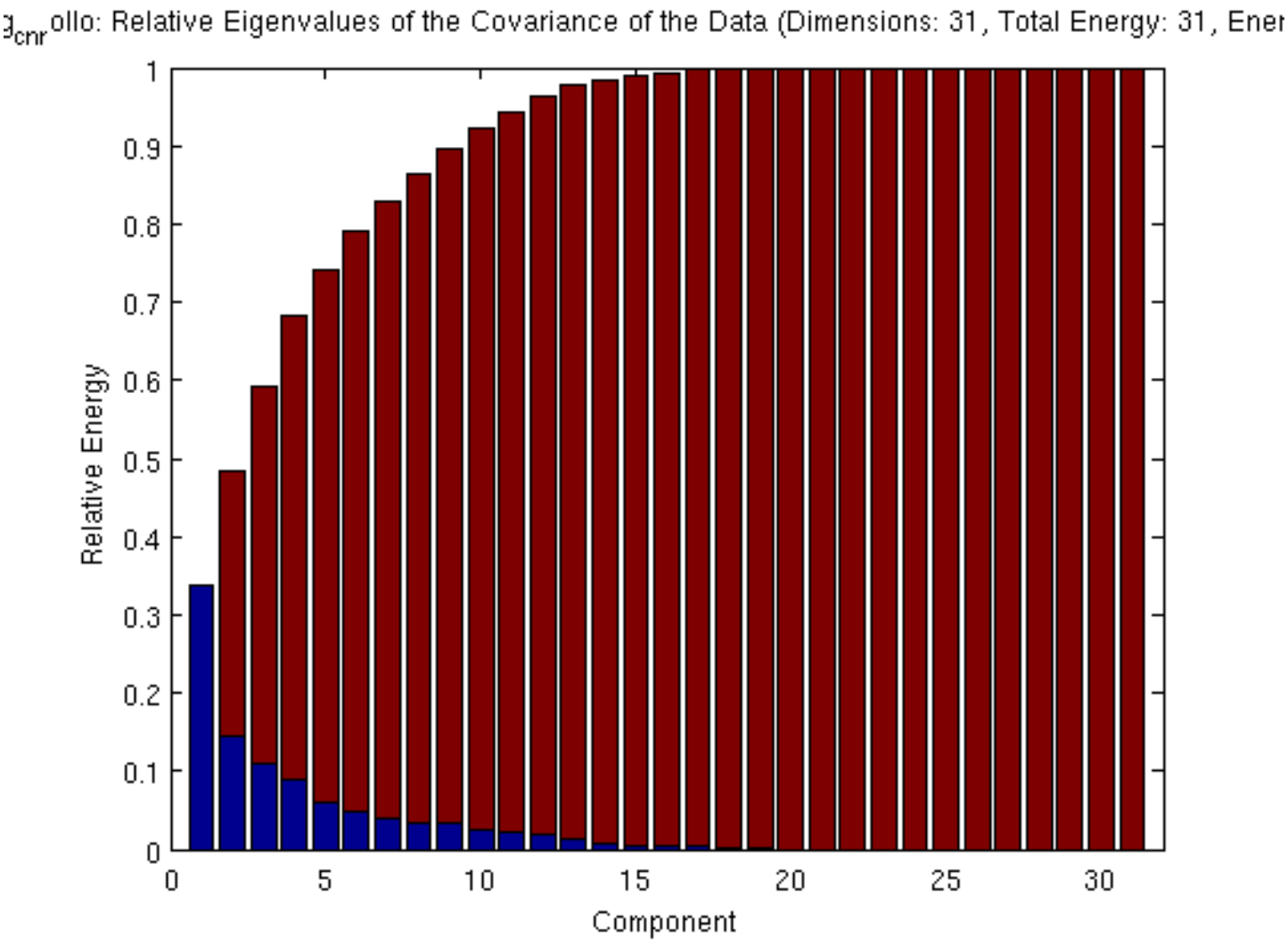
Dimension reduction

First of all, we need to know the amount of information of each attribute, to do that we can use the Principal Component Analysis method (PCA)

PCA

PCA is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance.

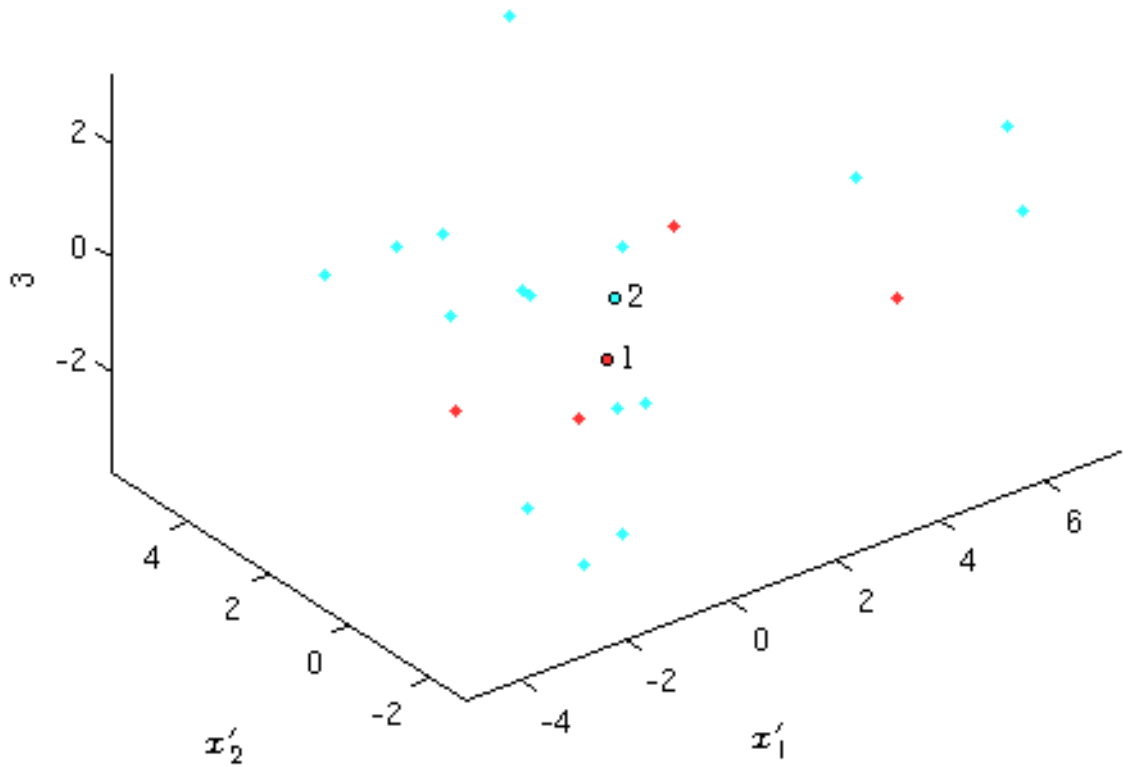
PCA results



Only with 3 dimension we maintain 60% of the information

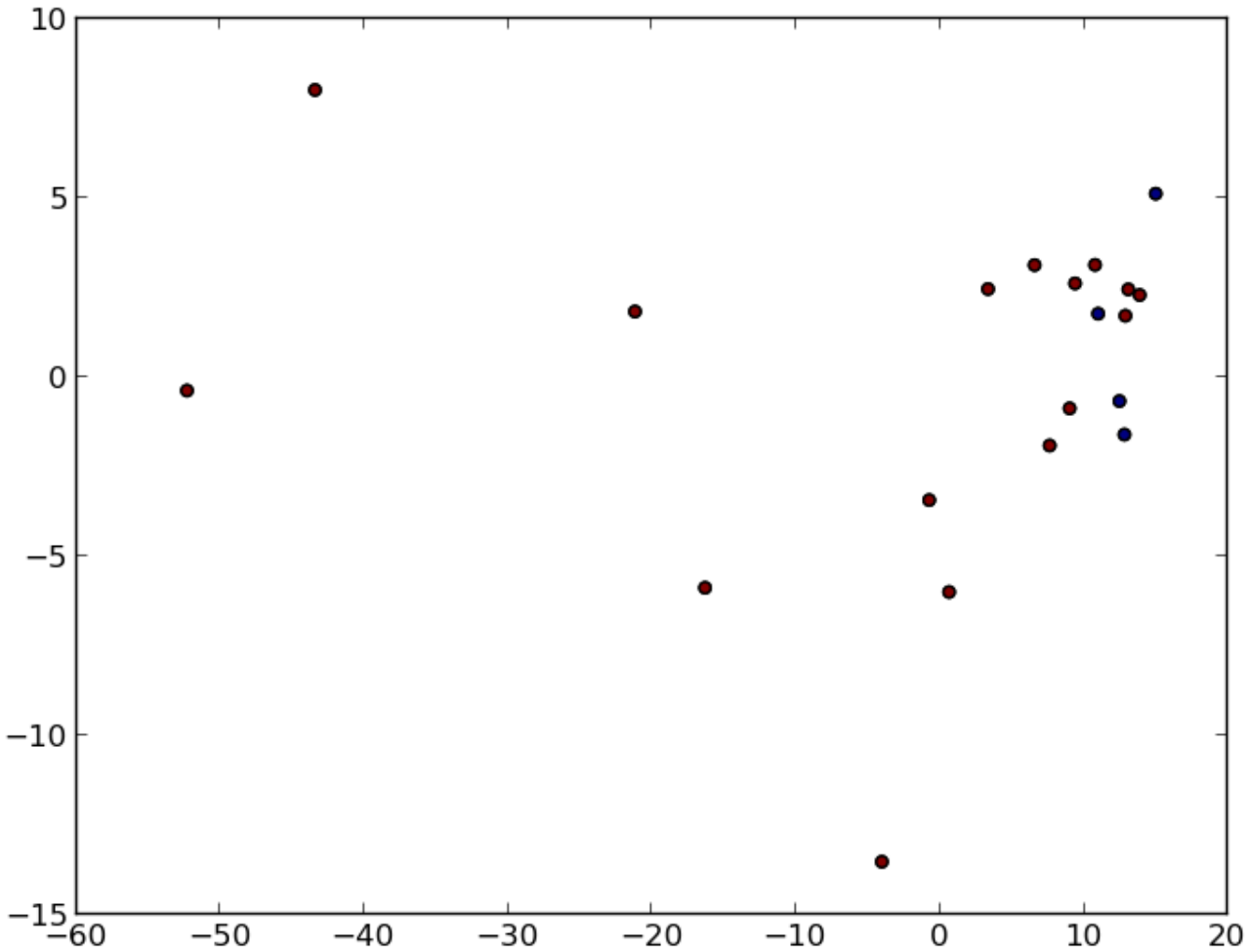
3D PCA plotting

$45, n = 5000, p_0 = 1/5 = 0.2, |\mathcal{Q}| = 2, k_{min} = 1000, p_{min} = 0.2(0\%), p_{max} = 0.8, \mathcal{D} = 1$



Positive points (red) are in the center of the space mixed with negative ones!!!

2D PCA plotting



Positive points (blue) are on a concrete area of the space

Manage the problem as an Anomaly Detection Task

- Manage *certified* instances as anomalies (only 4 cases)
- Evaluate if the anomalies are separable from the rest of the dataset
- Employ two different approaches:
  - + Unsupervised Anomaly Detection (LOF)
  - + Semi-Supervised Anomaly Detection (OSvm)

Unsupervised vs Semi-Supervised

Unsupervised

These methods do not assume any information about class label and use only distance based techniques to rank the instances in order to have in the top part of the rank the anomalies

Semi-supervised

These approaches exploit label information about the *normal* class. We can first train a model (only with instances coming from one class) and then perform a classification (or ranking) in order to detect instances far from the induced model.

Area Under the Curve

- Evaluate the quality of a learning model
- Compare TPR (True Positive Rate) with FPR (False Positive Rate)
- Compute the Area under the Curve varying the number of instances taken into consideration
- In our case compute TPR and FRP for the top-K objects in the ranking, varying the value of K

Results

Preliminary results using the value K=3 (nearest neighbors) for LOF and using a 5-fold cross validation for OSVM

Data	**Original**	**PCA** (3dim)
OSVM	0.4167	**0.4833**
LOF	0.3438	0.29

Table: Results from Anomaly Detection Algorithms

Considerations

- Results not so good for the moment
- Exploit label information can help the process, also partial as in this case
- Best results OSVM+PCA while LOF performs very bad
- Euclidean distance (LOF) is not adapt to tackle with this data while more sophisticated measure as radial based kernel (OSVM) can be better

Some conclusions

- We cannot learn but we can discard part of the negative instances easily
- We only need 2 or 3 dimensions at most
- Decision trees, k-NN or clustering are not suitable
- After some test SVM does not seem very promising too
- Case based reasoning or learning by example methods could work fine, if we obtain more positive examples
- Problem!!! $\rightarrow$ most of the instances only has 4 valued attributes instead of 31

Future work

Create lots of models

- As we have a file with all the components, generate an extensive collection of classifiers with WEKA
- Then, apply the guided search methodology
- and/or the model consensus methodology

Classifiers for rare events

- There is a library in R for predicting rare events: http://www.liaad.up.pt/ ltorgo/Research/
- 4 C instances over 97,718 instances are very rare events ...

Guided search (model disambiguation)

- Since we do not have enough information to decide which model is better, predict the rest of the rows with all the models.
- Compute the minimum set of observations that will allow deciding which model is better (set with maximum divergence of predictions between models).
- Inform astronomers of them so that we can have more informative results quickly.
- Once new data is available, retrain all models and check if there is some improvement.
- Iterate until a model reaches a desired classification accuracy or no progress is shown in any model.

Model consensus

- This is the dual strategy to Guided Search.
- Compute the set of predictions for which most models agree on the predictions. The idea is that a prediction error found in these predictions will have an impact on most of the classifiers. Positive predictions with consensus might be already of interest!
- Inform astronomers of them so that we can have the real values of the predictions.
- Once new data is available, retrain all models and check if there is some improvement.
- Iterate until a model reaches a desired classification accuracy or no progress is shown in any model.

Amadeus Exoplanete Data Analysis

Jordi Nin, Marc Solé, Dino Ienco

nin,msole@ac.upc.edu
dino.ienco@teledetection.fr

LIRMM-CNRS and DAC-UPC