Yeast Cells Classification

Machine Learning Approach to Discriminate Saccharomyces cerevisiae Yeast Cells Using Sophisticated Image Features.

Mohamed Tleis

Supervisor: Fons J. Verbeek

Leiden University

October 15, 2015

IB-2015

11th Integrative BioInformatics International Symposium

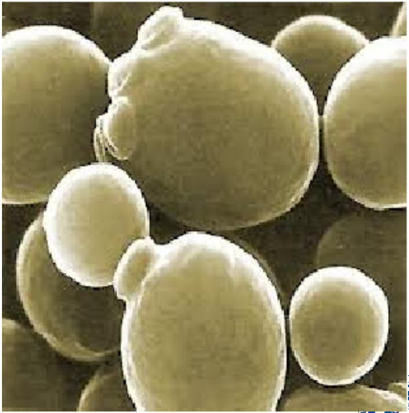
Table of Contents

Section 1

Introdution

Saccharomyces cerevisiae

- Originally isolated from grapes skin.
- Intensively studied eukaryotic model.
- Used to understand gene behaviour, under stress response.



Saccharomyces cerevisiae. .

BMH1-GFP under high stress 50mM *NaCl* media

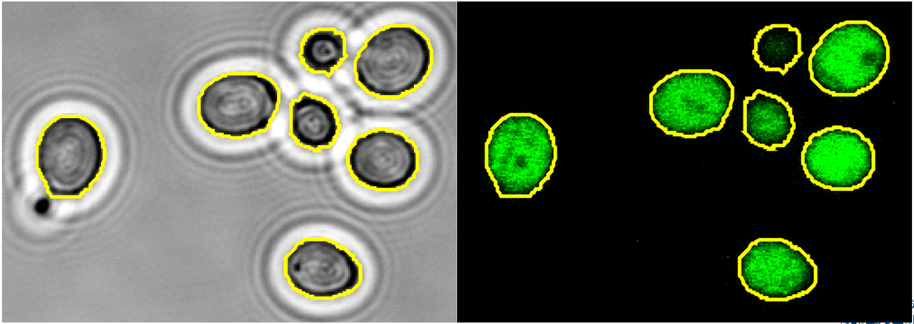


Image Analysis Platform

- Has the following components:
  - Segmentation Module.
  - Measurement Module.
  - Statistics and Data Visualization Module.
  - GUI.

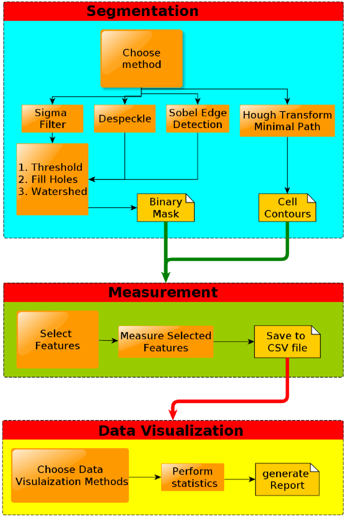
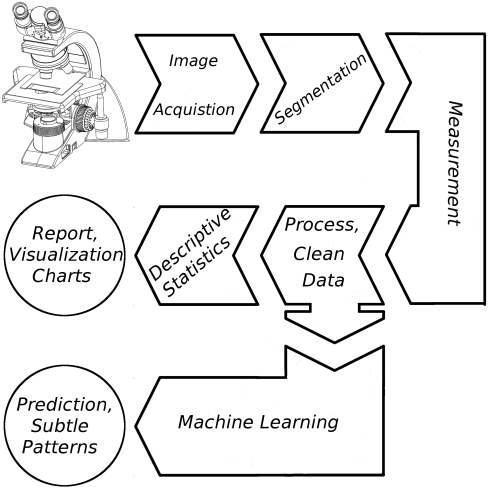
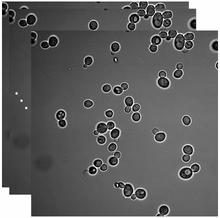


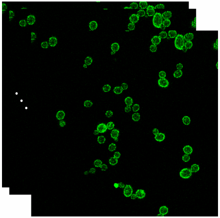
Image Analysis Workflow



Yeast Cells Image Modality

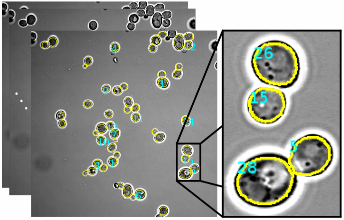
- Images Acquired by Zeiss LSM5 Exciter.
  - 2-Channels
    - Bright-Field
    - GFP- Protein

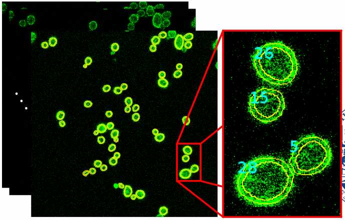




Yeast Cells Segmentation

- Segmentation on Bright-Field Channels.
- Resulted masks used to measure all channels.



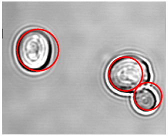


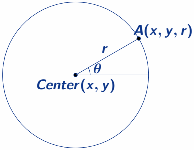
Hough Transform To Detect Circles

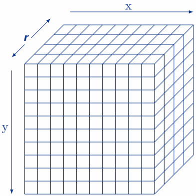
- Detect Geometrical Circles.
  - Using 3D cube-like Accumulator.
  - Threshold to estimate cell locations.

$$T = 2\pi r - \{2\pi r \times \alpha + p\}$$

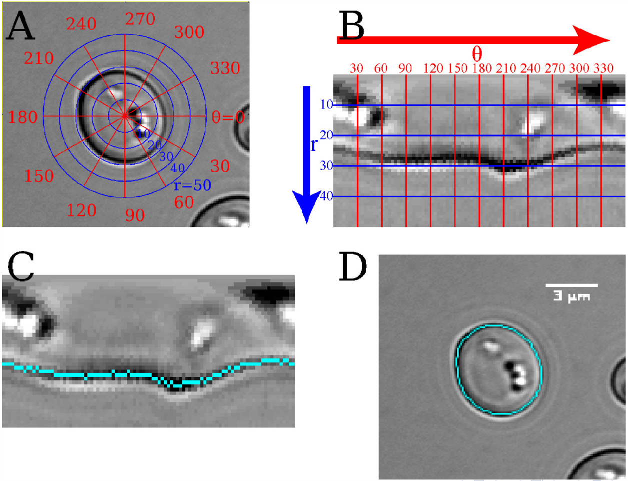
$$p = \beta \times (r_{max} - r_{min}) - r_{index}$$



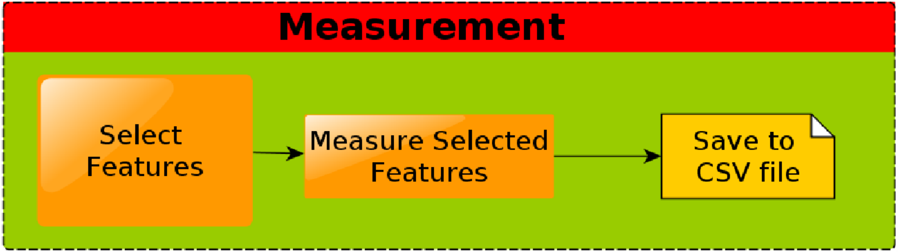




Dynamic Programming

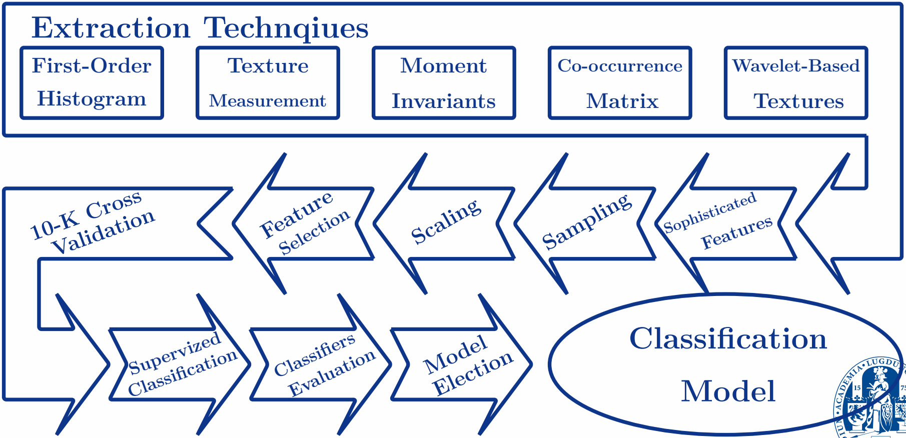


Measurement Module



- Subtle patterns not easy to be extracted.
- Sometimes it's not possible to see differences in cell groups.
- We need an automatic system to extract hidden features.

Machine Learning



Section 2

Features

Feature, Texture & Extraction Techniques

- A feature is a representation/attribute of an image.
- Texture :
  - is the visual effect produced by spatial distribution of variations.
  - is a rich Source of visual Information.
- Feature Extraction is locating pixels with distinctive characteristics.

First-Order Histogram

- An Image as a function $f(x, y)$.

$$h(i) = \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} \delta(f(x, y), i), \tag{1}$$

$$\delta(j, i) = \begin{cases} 1, j = i \\ 0, j \neq i \end{cases} \tag{2}$$

Features based on First-Order Histogram

Features	Description
Size	The number of pixels occupied by the cell.
Total Intensity	Sum of intensity values of pixels occupied by the cell.
Intensity Standard Deviation	The standard deviation from the mean (intensity/pixel) of the intensity values at each pixel.
Perimeter	Cell perimeter.
Circularity	The circularity of detected shapes. $$\mathit{Circularity} = \frac{4\pi \mathit{Size}}{\mathit{perimeter}^2}. \quad (3)$$
Vacuole Size	Estimation of the vacuole size.
Membrane Features	Size, total Intensity, Intensity standard deviation.

Textures based on First-Order Histogram

Textures	Description
Variance	Measure of intensity contrast. $\mu_2(z) = \sum_{i=0}^{L-1}(z_i - m)^2.P(z_i)$
Relative Smoothness	Zero for constant intensities. $R(z) = 1 - \frac{1}{1+\sigma^2(z)}$
Skewness	Indication of the skewness of the histogram. $\mu_3(z) = \sum_{i=0}^{L-1}(z_i - m)^3.P(z_i)$
Uniformity	Has a maximum value when intensity levels are equal. $U(z) = \sum_{i=0}^{L-1} P^2(z_i)$
Entropy	A measure of variability, is zero for constant images. $e(z) = -\sum_{i=0}^{L-1} P(z_i).log_2 P(z_i)$

Moment Invariants

- An image moment:
  - Is a certain particular weighted average, i.e. moment of pixel intensities.
  - Computed based on the information from shape and interior region.
  - Useful descriptor after segmentation.
- Simple Properties from low order moments.
- Invariant to translation, scale and rotation.
- frequently used as features for:
  - Image Processing.
  - Remote sensing.
  - Shape recognition.
  - Classification.

Hu's Set of Moment Invariants

$$hu = \{\Phi_1, \Phi_2, \Phi_3, \Phi_4, \Phi_5, \Phi_6, \Phi_7\}.$$

- Are widely Known set of seven invariants.
- Φ_1 & Φ_2 are based on second order moments.
- Φ_3 ... Φ_7 are based on third order moments.
- More effective when fused with other techniques.

Co-occurrence Matrix

- Simple texture attributes can not characterize cells.
- Similar texutres agree in their second-order statistics.
- Second Order statistics:
  - Are given by pairs of pixels.
  - Have good discrimination rates.
  - Important in automated image analysis.
  - Features derived from co-occurrence matrix.
- Co-occurrence Matrix:
  - For Image f(x,y) with L discrete levels (Dimension L x L).
  - The $\boldsymbol{(i,j)^{th}}$ is **#** of times that f(x1,y1) = i and f(x2,y2) = j.
  - where (x2,y2) = (x1,y1) + (d·cos $\boldsymbol{\theta}$, d·sin $\boldsymbol{\theta}$).

Co-occurrence Matrix Derived Features

- Features used for texture discrimination.
  - Angular Second Moment.
  - Correlation.
  - Intertia.
  - Absolute Value.
  - Inverse Difference.
  - Entropy.
  - Maximum Probability.

Multi-Scale Features

- Methods to calculate multi-scale features:
    - Wigner Ditributions.
        - Has interference terms between components.
    - Gabor Transform.
        - Non-orthogonal –> Redundant features.
    - Wavelet Transforms.

Section 3

Classification

Classification

- Dataset of 1440 yeast cell instances.
- 14-3-3 proteins with GFP in 50mM vs. 0mM $NaCl$
- Measure all features per individual cell instance
- Construct a contigency table to represent dispositions of the set of instances.
- Evaluate 23 different linear and non-linear classifiers.
  - ... including: decision trees, naive Bayes, least-square linear preictors, SVM, etc...

Imbalanced Dataset & Sampling Techniques

- Unequal distribution between classes.
- Sampling Techniques improves classifier accuracy.
  - UnderSampling.
  - OverSampling.
  - *SMOTE.*

Data Scaling, i.e. Normalization

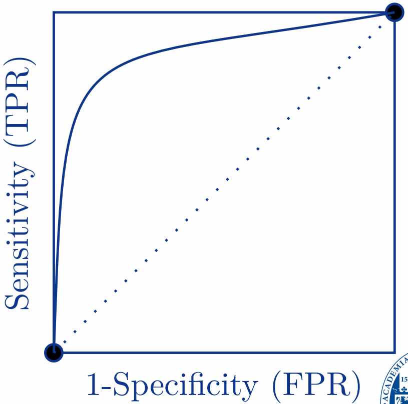
- Applied at data pre-processing.
- Some Algorithms will not work without Normalization.
- Normalization Techniques:
  - UL. $x_i{}^* = \frac{x_i}{||x||}, i = 1, 2, ...d,$
  - MV. $x_i{}^* = \frac{(x_i - \mu)}{\sigma}, i = 1, 2, ., d,$

Feature (Attribute) Selection

- To optimally reduce feature space.
- Advantages:
  - Improves the prediction performance.
  - Provides faster and more cost effective classifiers.
  - Provides a better understanding of the underlying process that generated the data.
  - Reduces overfitting.
  - Reduces training time.
- Avoid selecting redundant and irrelevant features.
- Selected Algorithms:
  - Information Gain (IG).
  - Correlation Feature Selection (CFS).
  - Principal Component Analysis (PCA).

Evaluation metrics, *ROC* and *AUC*

- *ROC* curve is a 2D graphical plot.
  - **$AUC = 1$**, Perfect
  - **$1 > AUC \geq 0.9$**, Excellent.
  - **$0.9 > AUC \geq 0.8$**, Good.
  - **$0.8 > AUC \geq 0.7$**, Fair.
  - **$0.7 > AUC \geq 0.6$**, Poor.



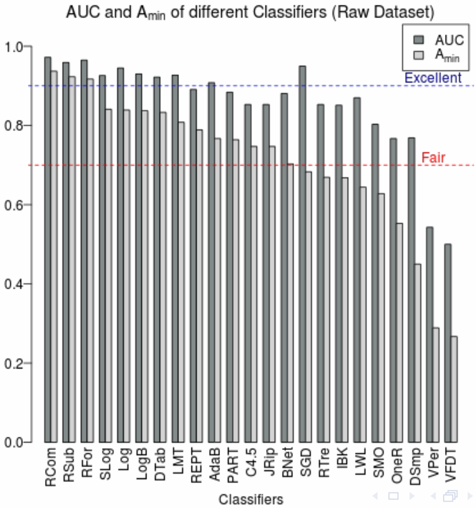
Classifiers Evaluated

- 23 different classifiers.
- Using *Weka*, *R* and *rWeka*

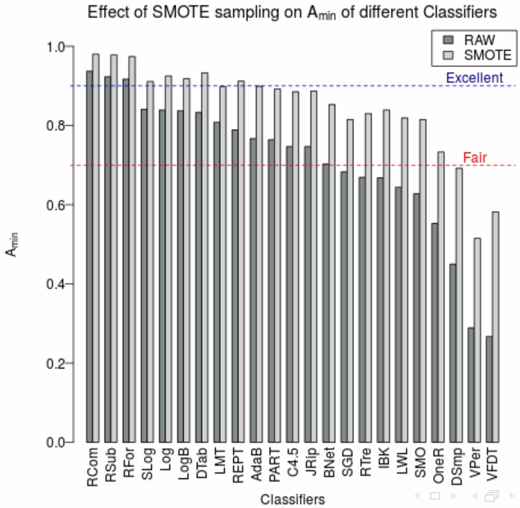
Section 4

Results

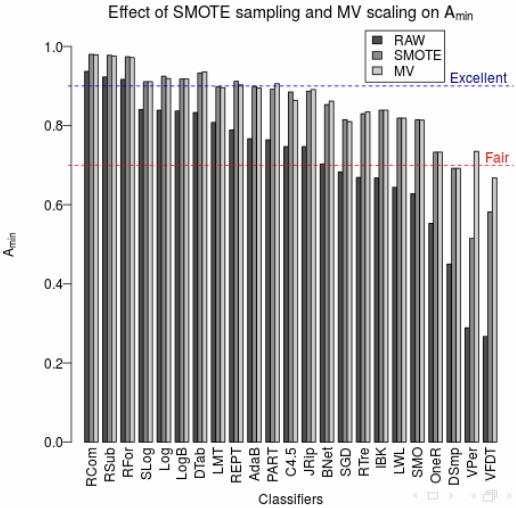
AUC vs. A_{min}


AUC and A_min of different Classifiers (Raw Dataset)
AUC
A_min
Excellent
Fair
1.0
0.8
0.6
0.4
0.2
0.0
RCom
RSub
RFor
SLog
Log
LogB
DTab
LMT
REPT
AdaB
PART
C4.5
JRip
BNet
SGD
RTre
IBK
LWL
SMO
OneR
DSmp
VPer
VFDT
Classifiers

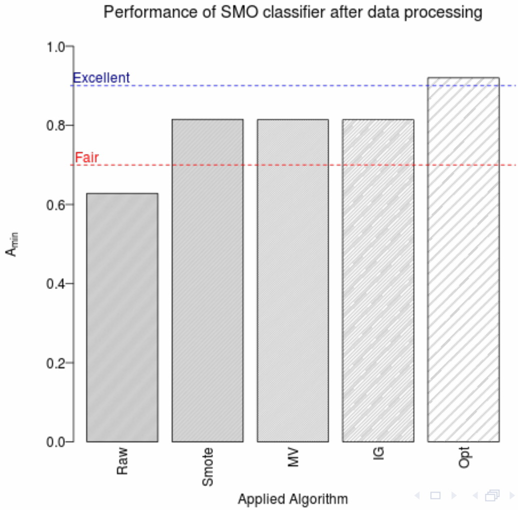
Power of Sampling



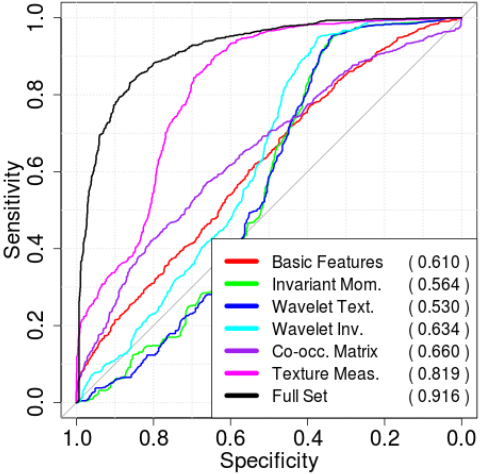
Normalization and Feature Selection



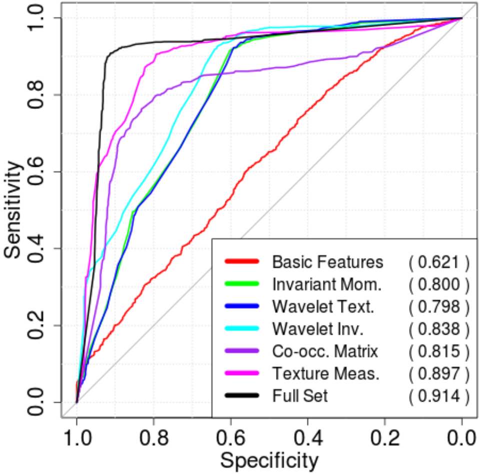
SVM : SMO


Performance of SMO classifier after data processing
1.0
0.8
0.6
0.4
0.2
0.0
Excellent
Fair
A_{min}
Raw
Smote
MV
IG
Opt
Applied Algorithm

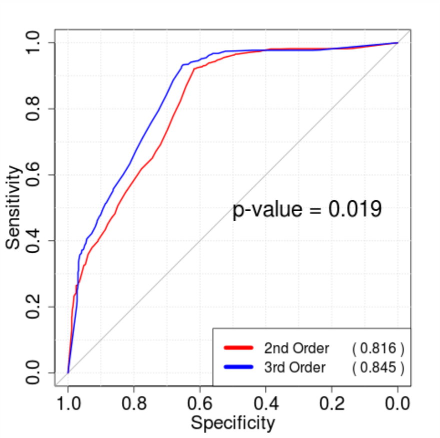
Analysis and *AUC* value of *Logistic* Classifier



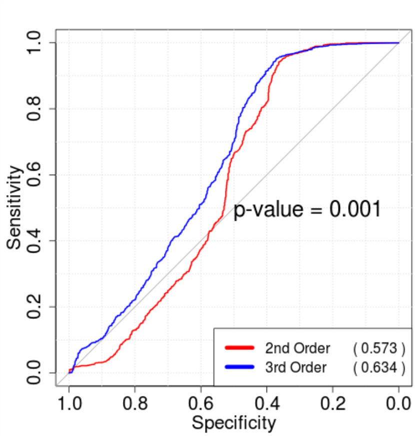
Analysis and *AUC* value of *C4.5* Classifier



Performance of Classifiers using second and up-to third order invariant moment features



C4.5



Logistic

Section 5

Conclusion

Conclusion

- A machine learning approach can discriminate yeast cells cultivated under different stress levels.
- A feature set is powerful in predicting cell groups, combined features from 1st-order histogram, moment invariants, Co-occurrence matrix and Wavelet-based texture features.
- Using *SMOTE* for data sampling, *MV* for data normalization and *IG* for feature selection.
- As future work:
  - Classify different cell strains and conditions in a high-volume HTS studies.
  - Use developmental techniques to create optimal classifier.

Acknowledgement

- Correspondence : {m.tleis, f.j.verbeek}@liacs.leidenuniv.nl

- Supervisor : Dr. Ir. Fons J. Verbeek
  (section Imaging **&** BioInformatics, LIACS)

- Contributors: