Yeast Cells Classification

Machine Learning Approach to Discriminate \textit{Saccharomyces cerevisiae} Yeast Cells Using Sophisticated Image Features.

Mohamed Tleis

Supervisor: Fons J. Verbeek

Leiden University

October 15, 2015

IB-2015

11\textsuperscript{th} Integrative BioInformatics International Symposium
Table of Contents

1. Introduction
2. Features
3. Classification
4. Results
5. Conclusion
Section 1

Introduction
Saccharomyces cerevisiae

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

- Image Acquisition
- Segmentation
- Measurement
- Descriptive Statistics
- Process, Clean Data
- Machine Learning
- Prediction, Subtle Patterns
- Report, Visualization Charts

Mohamed Tleis (Leiden University)
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.

\[ T = 2\pi r - \{2\pi r \times \alpha + p\} \]
\[ p = \beta \times (r_{\text{max}} - r_{\text{min}}) - r_{\text{index}} \]
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

Extraction Techniques
- First-Order Histogram
- Texture Measurement
- Moment Invariants
- Co-occurrence Matrix
- Wavelet-Based Textures

10-K Cross Validation
- Feature Selection
- Scaling
- Sampling
- Sophisticated Features

Supervised Classification
- Classifiers Evaluation
- Model Election

Classification Model

Mohamed Tleis (Leiden University) Yeast Cells Classification 12 / 40
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 $f(x, y)$.

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

$$\delta(j, i) = \begin{cases} 1, & j = i \\ 0, & j \neq i \end{cases} \quad (2)$$
# Features based on First-Order Histogram

<table>
<thead>
<tr>
<th>Features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>The number of pixels occupied by the cell.</td>
</tr>
<tr>
<td>Total Intensity</td>
<td>Sum of intensity values of pixels occupied by the cell.</td>
</tr>
<tr>
<td>Intensity Standard Deviation</td>
<td>The standard deviation from the mean (intensity/pixel) of the intensity values at each pixel.</td>
</tr>
<tr>
<td>Perimeter</td>
<td>Cell perimeter.</td>
</tr>
<tr>
<td>Circularity</td>
<td>The circularity of detected shapes.</td>
</tr>
</tbody>
</table>

\[
Circularity = \frac{4\pi \text{Size}}{\text{perimeter}^2}.
\]  

(3)

| Vacuole Size                  | Estimation of the vacuole size.                                            |
| Membrane Features             | Size, total Intensity, Intensity standard deviation.                        |
# Textures based on First-Order Histogram

<table>
<thead>
<tr>
<th>Textures</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance</td>
<td>Measure of intensity contrast.</td>
</tr>
<tr>
<td></td>
<td>( \mu_2(z) = \sum_{i=0}^{L-1} (z_i - m)^2 \cdot P(z_i) )</td>
</tr>
<tr>
<td>Relative Smoothness</td>
<td>Zero for constant intensities.</td>
</tr>
<tr>
<td></td>
<td>( R(z) = 1 - \frac{1}{1+\sigma^2(z)} )</td>
</tr>
<tr>
<td>Skewness</td>
<td>Indication of the skewness of the histogram.</td>
</tr>
<tr>
<td></td>
<td>( \mu_3(z) = \sum_{i=0}^{L-1} (z_i - m)^3 \cdot P(z_i) )</td>
</tr>
<tr>
<td>Uniformity</td>
<td>Has a maximum value when intensity levels are equal.</td>
</tr>
<tr>
<td></td>
<td>( U(z) = \sum_{i=0}^{L-1} P^2(z_i) )</td>
</tr>
<tr>
<td>Entropy</td>
<td>A measure of variability, is zero for constant images.</td>
</tr>
<tr>
<td></td>
<td>( e(z) = - \sum_{i=0}^{L-1} P(z_i) \cdot \log_2 P(z_i) )</td>
</tr>
</tbody>
</table>
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.
- \( \Phi_1 \) & \( \Phi_2 \) are based on second order moments.
- \( \Phi_3 \) ... \( \Phi_7 \) are based on third order moments.
- More effective when fused with other techniques.
Co-occurrence Matrix

- Simple texture attributes cannot characterize cells.

- Similar textures 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 $L \times L$).
  - The $(i,j)^{th}$ is $\#$ of times that $f(x_1,y_1) = i$ and $f(x_2,y_2) = j$.
  - Where $(x_2,y_2) = (x_1,y_1) + (d \cdot \cos \theta, d \cdot \sin \theta)$.
Co-occurrence Matrix Derived Features

- Features used for texture discrimination.
  - Angular Second Moment.
  - Correlation.
  - Intertia.
  - Absolute Value.
  - Inverse Difference.
  - Entropy.
  - Maximum Probability.
Method to calculate multi-scale features:

- Wigner Distributions.
  - Has interference terms between components.

- Gabor Transform.
  - Non-orthogonal \(\rightarrow\) 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 contingency 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 predictors, 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. \( x_i^* = \frac{x_i}{\|x\|}, \ i = 1, 2, \ldots, d, \)

- MV. \( x_i^* = \frac{(x_i - \mu)}{\sigma}, \ i = 1, 2, \ldots, d, \)
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.
  - $\text{AUC} = 1$, Perfect
  - $1 > \text{AUC} \geq 0.9$, Excellent.
  - $0.9 > \text{AUC} \geq 0.8$, Good.
  - $0.8 > \text{AUC} \geq 0.7$, Fair.
  - $0.7 > \text{AUC} \geq 0.6$, Poor.
Classifiers Evaluated

- 23 different classifiers.
- Using *Weka*, *R* and *rWeka*
Section 4

Results
AUC vs. $A_{\text{min}}$
Power of Sampling

Effect of SMOTE sampling on $A_{min}$ of different Classifiers

- RAW
- SMOTE

Excellent

Fair

Classifiers:
- RCom
- RSub
- RFor
- SLog
- Log3
- LogB
- Dtab
- LMT
- REPT
- AdaB
- PART
- C4.5
- JRip
- BNet
- SGD
- RTre
- IBK
- LWL
- SMO
- OneR
- DSmp
- VPer
- VDOT

Mohamed Tleis (Leiden University)

Yeast Cells Classification
Normalization and Feature Selection

Effect of SMOTE sampling and MV scaling on $A_{\min}$

Classifiers: RCom, RSub, RFor, SLog, LogB, LogD, LMT, REPT, AdaB, PART, C4.5, JRip, BNet, SGD, RTre, IBK, LWL, SMO, OneR, DSm, VPer, VFDT.
SVM : SMO

Performance of SMO classifier after data processing

- Excellent
- Fair

A_min

Applied Algorithm:
- Raw
- SMO
- MV
- IG
- Opt
Analysis and AUC value of Logistic Classifier
Analysis and **AUC** value of **C4.5 Classifier**

![AUC Graph]

- **Basic Features**: 0.621
- **Invariant Mom.**: 0.800
- **Wavelet Text.**: 0.798
- **Wavelet Inv.**: 0.838
- **Co-occ. Matrix**: 0.815
- **Texture Meas.**: 0.897
- **Full Set**: 0.914
Performance of Classifiers using second and up-to third order invariant moment features

C4.5

Logistic

Mohamed Tleis (Leiden University)  Yeast Cells Classification
Section 5

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: