

Classification Methods for the Analysis of High Content Screening Data

Rich Jacques

Department of Probability & Statistics
University of Sheffield

11th October 2007



Outline

- Introduction
- Existing Methodologies
- Objectives and key problems
- New classification updating algorithm
- Model Convergence
- Conclusions

2

Introduction

- In drug discovery the aim is to find compounds to fight a particular disease
- High Throughput Screens (HTS) allow a large number of compounds to be tested in a biological assay
- HTS produce high content images, one for each compound
- Advanced imaging algorithms used to produce a set of variables for each image

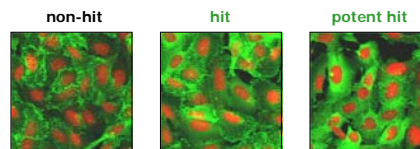


3

Supervised Classification Problem

- We wish to classify compounds into groups of hits and non-hits

Example Images:



4

Supervised Classification Problem

Data:

- 15 measurements (variables) are taken for each compound
- Data is split into three batches
- Batch 1 contains 12,285 compounds selected because of their known properties
- Batch A and B contain 33,941 and 33,408 compounds respectively

5

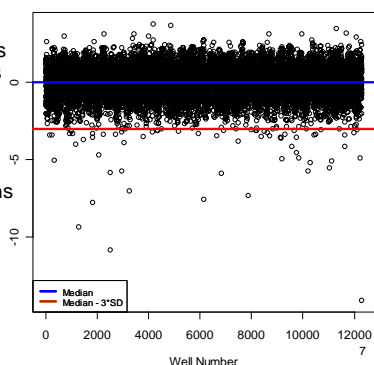
Current Methodology

- Compounds are classified using a single parameter
- All compounds classified as hits have their images checked by eye & classified as **true hits** or **false hits**
- Compounds not classified as hits are not checked by eye

6

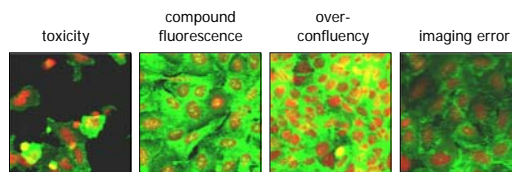
Current Methodology

- Outlying compounds are identified as hits
- Compounds are selected that differ from the mean by c standard deviations
- $c = 3$ used here



Current Methodology

Examples of false hits:



8

Objectives

- Develop a multiparametric approach to enable a more refined compound selection
- Reduce the number of false positives in order to reduce the 'cost' of manual image inspection
- Avoid an increase in the number of false negatives since these are potential new drugs

9

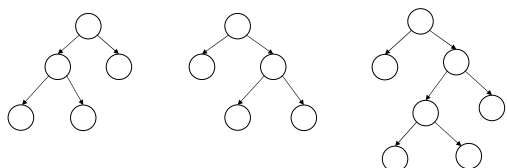
Preliminary Analysis

- Existing multivariate methodologies were used to classify compounds
- Batch 1 is used to train the classifiers and batches A and B are used as test data

10

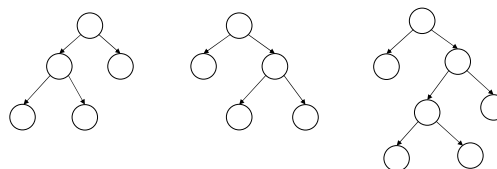
Random Forests

- A Random Forest is a classifier consisting of a collection of unpruned classification trees.
- Each tree is grown using a bootstrap sample of observations as the training data.
- The nodes on each tree represent the best split based on a random sample of m variables.



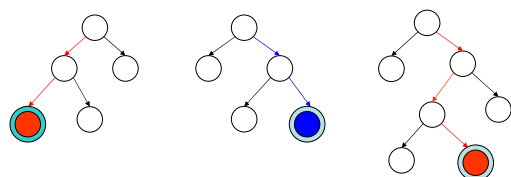
11

Random Forests



12

Random Forests



- Observations are classified by choosing the most frequently occurring of the classes as determined by the individual trees in the forest.

13

Preliminary Analysis

	1 Parameter	Linear Discriminant Analysis	Random Forest
Hit	69	73	96
Good Hit	51	50	49
Non-Hit	0	52	50
Focus Error	31	3	7
High Background	21	2	6
Over Confluent	5	14	45
Toxic	10	2	25
Well Dry	4	0	1
No Visible Image	10	0	0
Low Drag5	3	0	0
% Misclassified	41.18%	37.24%	48.03%

14

Preliminary Analysis

	1 Parameter	Linear Discriminant Analysis	Random Forest
Hit	69	73	96
Good Hit	51	50	49
Non-Hit	0	52	50
Focus Error	31	3	7
High Background	21	2	6
Over Confluent	5	14	45
Toxic	10	2	25
Well Dry	4	0	1
No Visible Image	10	0	0
Low Drag5	3	0	0
% Misclassified	41.18%	37.24%	48.03%

15

Preliminary Analysis

	1 Parameter	Linear Discriminant Analysis	Random Forest
Hit	69	73	96
Good Hit	51	50	49
Non-Hit	0	52	50
Focus Error	31	3	7
High Background	21	2	6
Over Confluent	5	14	45
Toxic	10	2	25
Well Dry	4	0	1
No Visible Image	10	0	0
Low Drag5	3	0	0
% Misclassified	41.18%	37.24%	48.03%

16

Key Problem

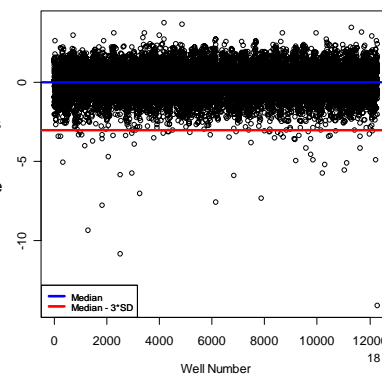
- Traditional multivariate classification methods assume the training data is a random sample from the **same distribution** as the test data
BUT
- Training data from HCS are selected because of their known biological effects
- We propose a method for updating classification rules as a solution to this problem

17

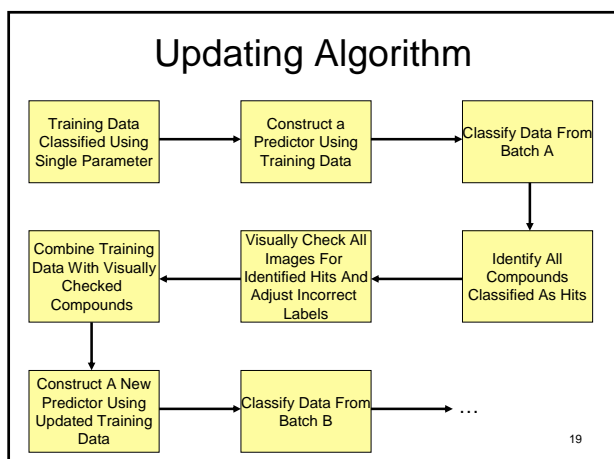
Updating Algorithm

Training Data Classified Using Single Parameter

- Outlying compounds are identified as hits
- Images of all hits are checked to ensure correct classification
- The images from a random sample of non-hits are also checked



Updating Algorithm



19

Comparing Methods of Discrimination

	1 Parameter	Linear Discriminant Analysis	Random Forest	Updated Random Forest
Hit	69	73	96	97
Good Hit	51	50	49	49
Non-Hit	0	52	50	1
Focus Error	31	3	7	0
High Background	21	2	6	0
Over Confluent	5	14	45	0
Toxic	10	2	25	0
Well Dry	4	0	1	0
No Visible Image	10	0	0	0
Low Drag5	3	0	0	0
% Misclassified	41.18%	37.24%	48.03%	0.68% ²⁰

Comparing Methods of Discrimination

	1 Parameter	Linear Discriminant Analysis	Random Forest	Updated Random Forest
Hit	69	73	96	97
Good Hit	51	50	49	49
Non-Hit	0	52	50	1
Focus Error	31	3	7	0
High Background	21	2	6	0
Over Confluent	5	14	45	0
Toxic	10	2	25	0
Well Dry	4	0	1	0
No Visible Image	10	0	0	0
Low Drag5	3	0	0	0
% Misclassified	41.18%	37.24%	48.03%	0.68%

Comparing Methods of Discrimination

	1 Parameter	Linear Discriminant Analysis	Random Forest	Updated Random Forest
Hit	69	73	96	97
Good Hit	51	50	49	49
Non-Hit	0	52	50	1
Focus Error	31	3	7	0
High Background	21	2	6	0
Over Confluent	5	14	45	0
Toxic	10	2	25	0
Well Dry	4	0	1	0
No Visible Image	10	0	0	0
Low Drag5	3	0	0	0
% Misclassified	41.18%	37.24%	48.03%	0.68%

Model Convergence

- 'Model convergence' is important with any interactive algorithm
- Does the 'model' converge to the same result if the ordering of the batches is changed?
- Each test batch split into 2 groups (A1, A2, B1, B2)
- Randomly assign order of batches in the algorithm
- 8 tests conducted

23

Model Convergence

Batch Order	Hit	Good Hit	Non Hit	Focus Error	High Background	Over Confluent	Toxic
A1, B1, A2, B2	95	52	1	2	0	3	1
A1, B2, A2, B1	96	50	1	1	0	0	0
A2, B1, B2, A1	94	49	3	1	0	2	0
A2, B2, A1, B1	91	50	2	0	0	0	0
B1, A2, B2, A1	83	48	2	0	0	0	0
B1, B2, A1, A2	84	49	2	1	3	0	0
B2, A1, A2, B1	87	49	3	0	0	0	0
B2, A1, B1, A2	83	49	2	0	1	0	0

24

Model Convergence

Batch Order	Hit	Good Hit	Non Hit	Focus Error	High Background	Over Confluent	Toxic
A1, B1, A2, B2	95	52	1	2	0	3	1
A1, B2, A2, B1	96	50	1	1	0	0	0
A2, B1, B2, A1	94	49	3	1	0	2	0
A2, B2, A1, B1	91	50	2	0	0	0	0
B1, A2, B2, A1	83	48	2	0	0	0	0
B1, B2, A1, A2	84	49	2	1	3	0	0
B2, A1, A2, B1	87	49	3	0	0	0	0
B2, A1, B1, A2	83	49	2	0	1	0	0

25

Conclusions

- Updated Random Forest has the smallest misclassification rate
 - One selected hit misclassified
 - (but does require intermediate visual check)
- Reordering batches leads to the same classifications (approx)
 - Small difference between groups of orderings starting with A and B

26

• Acknowledgements

- Edward Ainscow and Nick Brown
 - **Advanced Science and Technology Laboratory,**
AstraZeneca Charnwood
- Chris Harbron
 - **AstraZeneca Alderley Park**



27