COLLEGE OF SCIENCE
THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

Research Horizons Day & Research Week March 11-15, 2019

Research Horizons Day

Else-Tree Classifier for Minimizing Misclassification of Biological Data

Truong X. Tran, Computer Science Dept., Marc L. Pusey, IXpressGenes Inc., Ramazan S. Aygun Computer Science Dept.

Overview

In many applications, such as biological research, inaccuracy or misclassification of machine learning algorithms can yield fatal results. A significantly high value of an evaluation measure is an indication of overfitting, and chosen classifiers are likely to have false classifications for new data.

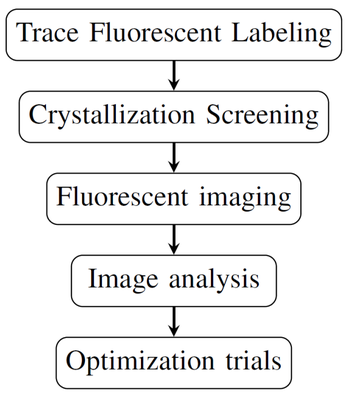
Else-Tree is a novel machine learning classifier that reduces the misclassification of data samples by labeling them as undecided rather than assigning them an incorrect class.

Impact

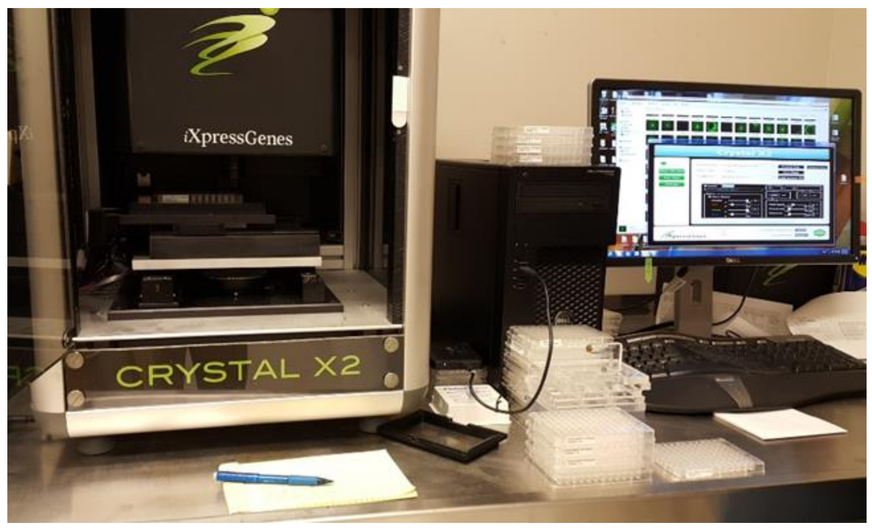
The Else-Tree both avoids critical mistakes and increases the trust of the user of the classifier.

Protein Crystallization Research:

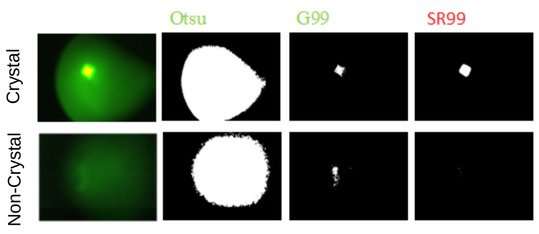
- Thousands of trials may need to be set up for a successful crystalline outcome.
- The 3-D structure of a protein is initially obtained by crystallizing the protein in drug development, missing a crystalline condition may hinder its development.
- By using Else-Tree, the expert may only review undecided items rather than all samples.



Stages of protein crystallization analysis using Trace Fluorescent Labeling.



Fluorescent microscope system, Crystal X2, for finding protein crystal leads.

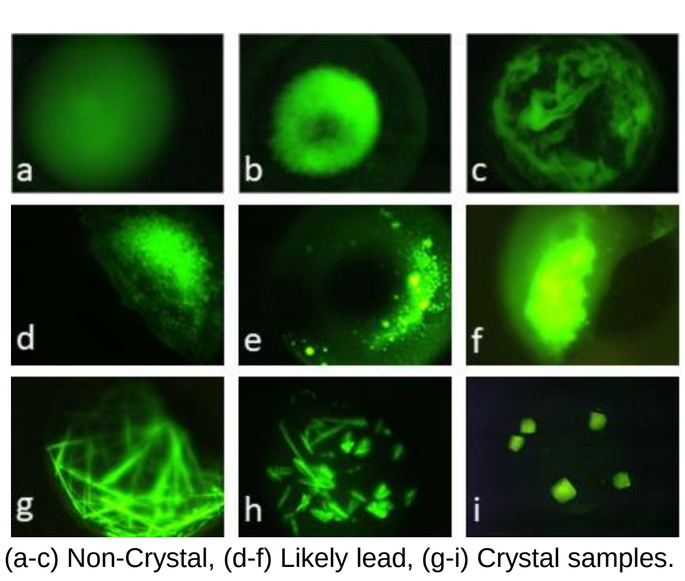


Left to right: original trial images, and binary images generated by different techniques

PROTEIN CRYSTALLIZATION TRIAL IMAGES FEATURE DATA DISTRIBUTION

	Training			Testing		
Data	1*	0*	Total	1*	0*	Total
T100	1156	1600	2756	1156	1600	2756
T90	1035	1445	2480	121	155	276
T80	930	1275	2205	226	325	551
T70	823	1106	1929	333	494	827

∗ 1: Likely-lead and Crystal, 0: Non-Crystal



(a-c) Non-Crystal, (d-f) Likely lead, (g-i) Crystal samples.

Other fields:

- Brest Cancer Dataset (BC), Banknote authentication (BN), Wireless Indoor Localization Dataset (WI).

Key Findings

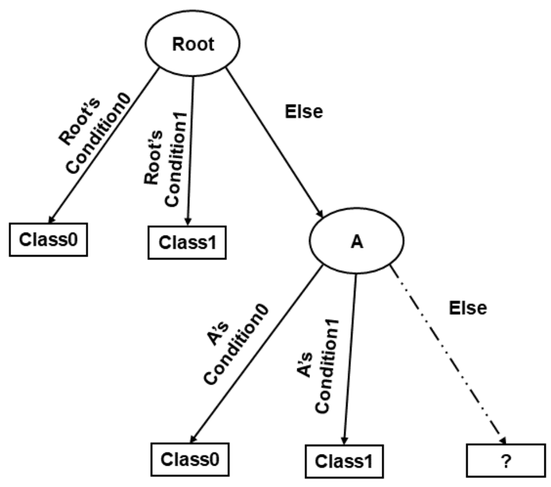
The main feature of the else-tree is its potential to generate zero percent error without overfitting by separating hard to classify data as undecided.

Explanation

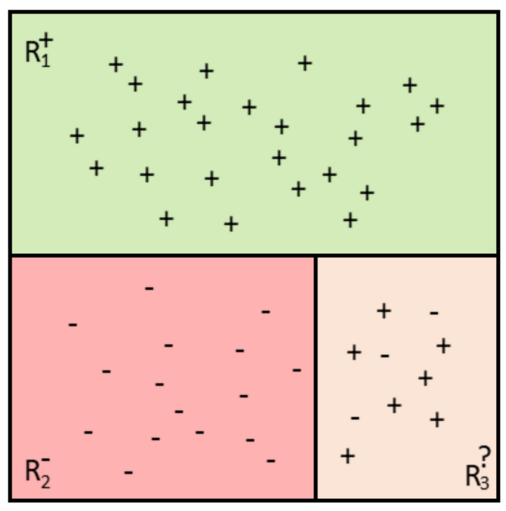
An Else-Tree classifier contains nodes, branch conditions, and leaf nodes having labels. The novelty of Else-Tree is its *else* branch and the final else branch leading to *else-leaf. Rather than* giving the wrong prediction, Else-Tree marks doubtful samples as '?' or undecided.

The Else-Tree is built by analyzing pure regions of an attribute per class of the training data. The most populated contiguous regions per class are used to label leaf nodes. The rest of the data ranges are fed into the *else* branch to recursively build the tree.

For classifying, if the new instance falls into a leaf node, the label of the leaf node is assigned. Otherwise, the else-branch is followed for using another attribute to classify. The last else-branch takes to the else-leaf. Any data that goes into this else-leaf is classified as undecided.



A sample of an Else-Tree structure for binary classification problem.



Rectangular decidable and undecided region.

Input (x_j, y):	2,0	4,1	1,0	5,1	6,1	4,0	2,0	3,1	3,0	1,0	6,1	5,1
Sort by x_j :	1,0	1,0	2,0	2,0	3,0	3,1	4,0	4,1	5,1	5,1	6,1	6,1
Region partition:	[1,2],0					[3.4],?				[5,6],1		

Example of partitioning dataset with respect to attribute j having values in the range of [1,6]. First, there were two classes of Y, as 0 and 1. Then, the non-uniform range is marked as undecided.

Experiment results

FALSE POSITIVE AND FALSE NEGATIVE RATE COMPARISON

	FP			FN			U
Dataset	C4.5	RF	ElseT	C4.5	RF	ElseT	ElseT
T100	0.7%	0%	0%	0.2%	0%	0%	14.55%
T90	0.6%	1.9%	0%	5.0%	0.8%	0%	16.67%
T80	1.8%	1.5%	0%	2.2%	0.4%	0%	14.51%
T70	1.8%	1.8%	0.4%	0.3%	0.3%	0%	12.09%

CORRECT CLASSIFICATION RATE IN 90% HOLD-OUT VALIDATION (LAST COLUMN IS FOR UNDECIDED RATE OF ELSE TREE)

Data	NB	SVM	MLP	C4.5	RF	ElseT	U
BC	94.74%	98.25%	94.74%	92.98%	94.74%	100%	10.52%
BN	85%	98.57	99.28%	98.57%	99.28%	100%	10.00%
WI	97.50%	99.50%	98.50%	98.00%	99.50%	100%	5.00%

Acknowledgements

This research was supported by National Institutes of Health (GM116283) grant and iXpressGenes, Inc.

This study was presented at MABM in 2018 IEEE International Conference on Bioinformatics and BioMedicine.

NATIONAL INSTITUTES OF HEALTH

iXpressGenes