Auto-Annotation of Cell Types

Midterm Project – STAT 454 / 556

Dayten J. Sheffar
Leno S. Rocha

Department of Mathematics and Statistics

March 2021

Real-world problem and its importance

- In analyzing single-cell RNA sequencing data, it is crucial to discover what cell types exist within a particular sample. This process is often the first step in advancing our knowledge of cellular dynamics [1], however, manual annotation is partially subject to misinterpretation [2] or other biases, and it can be very time consuming.
- Various methods have been developed to associate gene expression data of single cells to a cell type; one can access curated marker gene databases [3] [4], correlate the gene expression profiles between samples [2], and more recently, transfer labels with supervised classification.
- This work evaluates the performance of several classification models for cell type annotation using gene expression profiles and the disease status of patients as predictors.

https://doi.org/10.1016/j.csbj.2021.01.015 [1]
https://doi.org/10.1165/rcmb.2017-0430TR [2]
https://doi.org/10.1093/nar/gky900 [3]
https://academic.oup.com/nar/article/47/D1/D721/5115823 [4]

University of Victoria

Exploratory data analysis

Summary of variables in our dataframe:

1. Predictors: column `condt` [4954 obs.], containing a binary status marker for 'control' or 'ILD'; column `count`, a table with entries of type double [18339 x 4954 obs.].
2. Labels: `celltype` [4954 obs.] which has 5 classes of type character, as denoted in the table below.
3. `count`: only 120 thousand unique values in `count` (of over 91 million cells); a 0.13% density of input types. Over 81 million zeroes, & over 9 million positive continuous values, a 10% input density.
4. `count` has 18391 genes as column labels, e.g. 'FAM87B', and as sample labels there are 4954 6-character labels with various nucleobase sequences appended to them, such as 'F00409_AAGACCTCAAAGGTGC'.

Cell Type	*Control*	*ILD(sample%)*	**#c.typ.**	*0's in 'count'*	*ILD 'count' 0's*
AT1	472	299 (39%)	771	710 (92%)	279 (93%)
B Cells	140	857 (86%)	997	890 (89%)	771 (90%)
Basal	65	1993 (97%)	2058	1844 (90%)	1785 (90%)
Fibroblasts	140	287 (67%)	427	385 (90%)	258 (90%)
Mast Cells	254	447 (64%)	701	630 (90%)	403 (90%)
Total	1071	3883 (78%)	**4954**	4459 (90%)	3496 (90%)

University of Victoria

More EDA & Feature Selection

1. We removed 1,031 all-zero columns from `count` before conducting Kruskal-Wallis (H) one way ANOVA, a non parametric method, the statistic is found via

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

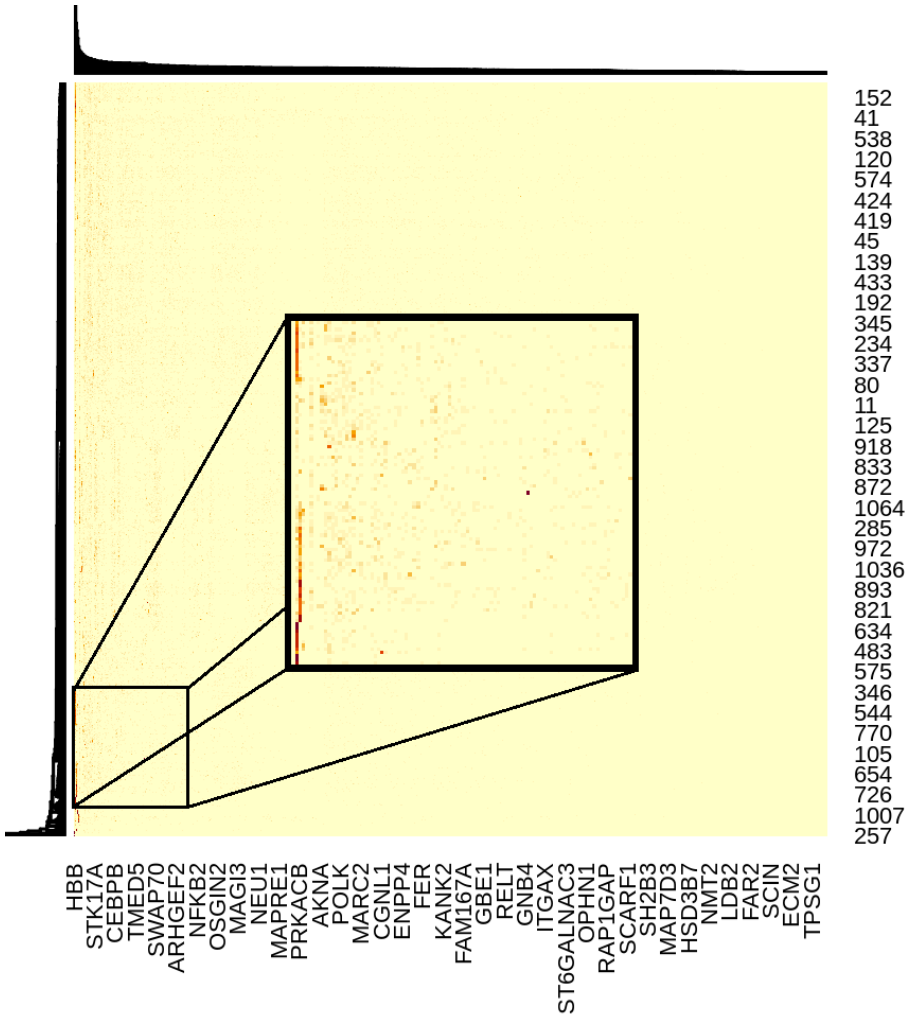
   where R_i is the sum of ranks of sample i, n_i is the size of sample i, N is the sum of sample sizes, and k is the number of samples – tied observations are averaged.

2. We arbitrarily used a p-value cutoff of 0.005, leaving 1,645 predictors.
3. Then adopted a correlation cutoff of 0.25, to avoid collinearity within the 1,645 predictors, leaving 1,065 predictors.

University of Victoria

Heat map - Selected features

Due to high dimensions and correlation cutoff, we zoom in on a smaller portion to demonstrate internal behaviour.



University of Victoria

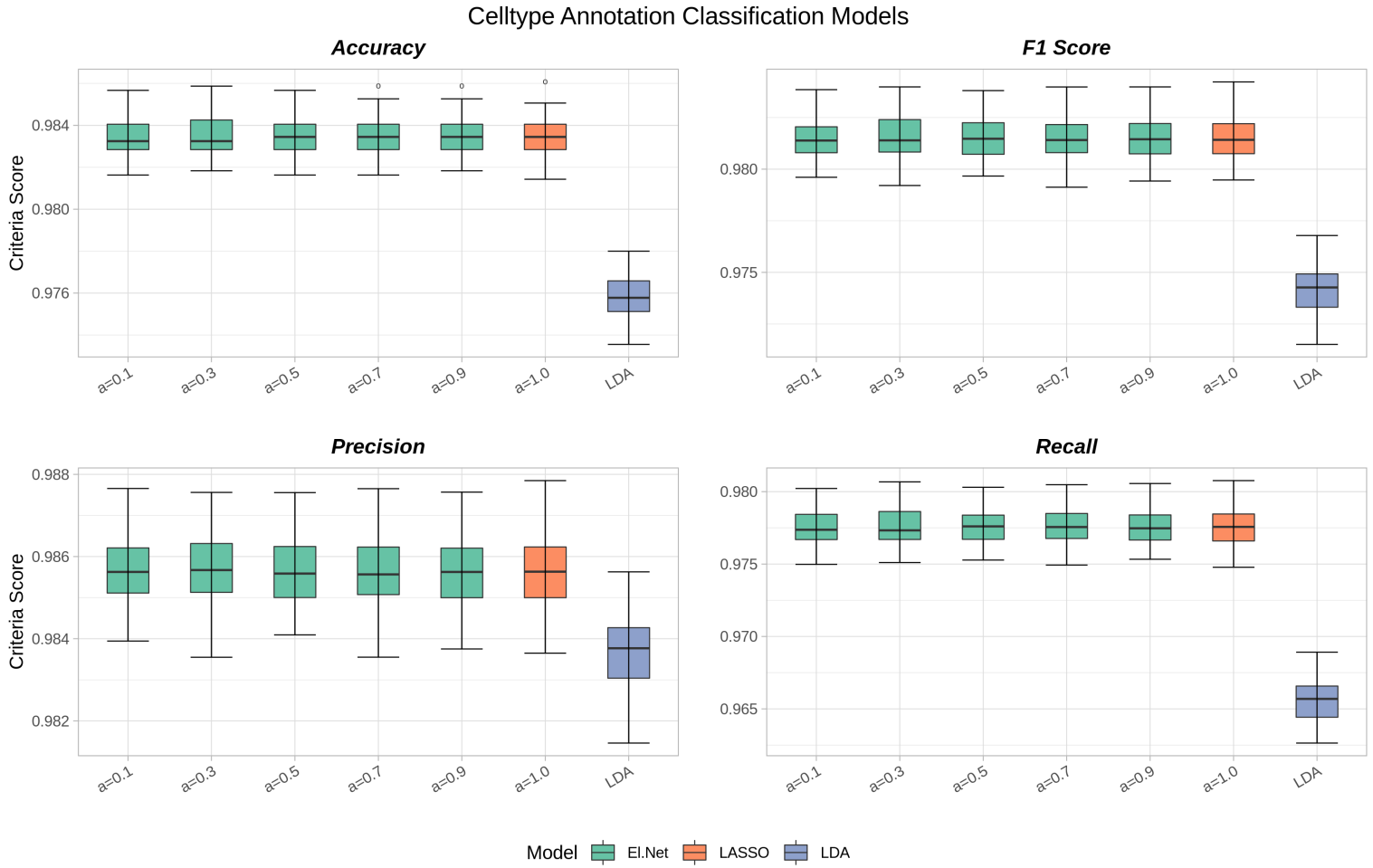
Methodology and Models

- Job submitted to Compute Canada for 10-fold Cross Validation 100 times.
- Time to run was 20.1 hours on 64 cores with 8Gb memory per cpu.
- In `R`, we ran the code in parallel using `foreach` and `doParallel` packages with file backed matrices from `bigstatsr` to fit the following models

1. Lasso (L1 penalized linear regression)
   `cv.glmnet(...,` α `= 1, family = 'multinomial')`
2. 5 Elastic Nets
   `cv.glmnet(...,` α `= c(0.1,0.3,0.5,0.7,0.9), family = 'multinomial')`
3. LDA (Linear Discriminant Analysis) – `MASS::lda()`
4. Classification Tree by `rpart()`, performing `rpart::prune()` to trim the tree with a complexity parameter chosen for parsimony, that is, for the simplest model (tied) within one standard error of the achieved minimum.

University of Victoria

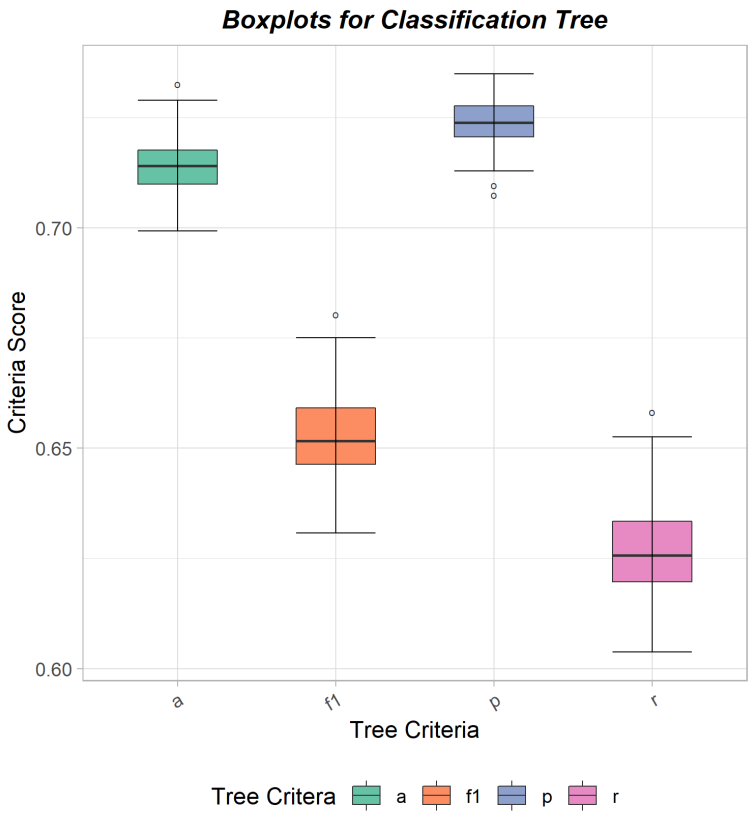
Visualization of Results



University of Victoria

CLTree left off due to poor performance, see next slide.

Results (*cont.*)



Boxplot Recap:

- Elastic net with $\alpha = 0.3$ generally performed best (on 3/4 metrics), scoring near 0.98 or above in all metrics.
- LDA underperformed Elastic Net and LASSO (the latter on par with Elastic Nets).
- Pruned classification tree significantly the worst fit accross all models and scores (0.6-0.7 on criteria).

University of Victoria

Performance metrics of the models

	Model	$\overline{Accuracy}$	p-value	Model	$\overline{Precision}$	p-value	Model	$\overline{Recall}$	p-value	Model	$\overline{F_1}$ score	p-value
1.	a=0.3	0.983496	–	a=0.3	0.985696	–	a=0.5	0.977623	NA	a=0.3	0.981525	–
2.	a=0.5	0.983488	9.19075e-01	a=0.9	0.985670	5.45509e-01	a=0.9	0.977597	6.41415e-01	a=0.5	0.981521	8.51871e-01
3.	a=0.9	0.983482	8.81943e-01	a=0.1	0.985669	7.05568e-01	a=0.3	0.977591	4.82887e-01	a=0.9	0.981516	9.14037e-01
4.	a=0.7	0.983474	4.53733e-01	a=0.5	0.985654	3.61394e-01	a=0.7	0.977577	4.79488e-01	a=0.7	0.981492	3.72198e-01
5.	a=0.1	0.983449	5.25935e-01	a=0.7	0.985643	1.29808e-01	a=1.0	0.977537	2.58360e-01	a=0.1	0.981482	5.17190e-01
6.	a=1.0	0.983447	4.71160e-01	a=1.0	0.985637	1.51089e-01	a=0.1	0.977532	1.73686e-01	a=1.0	0.981468	3.10247e-01
7.	LDA	0.975801	3.89281e-18	LDA	0.983682	3.58860e-17	LDA	0.965502	3.95591e-18	LDA	0.974115	3.95591e-18
8.	TREE	0.714184	3.94877e-18	TREE	0.724120	3.95591e-18	TREE	0.626545	3.95591e-18	TREE	0.652554	3.95591e-18

- The p-values were found between best model and other models, i.e., all compared to best model.
- Lasso and some Elastic Nets are not significantly better among themselves, but there is statistical evidence that they were better them LDA and Classification Tree.
- Tree has poor recall, bringing down its F_1 score.

University of Victoria

Conclusions

- The analysis of data from genetic cell expressions is currently a ‘hot’ research topic, with cell type annotation the starting point for many data analyses.
- We performed best subset selection of features by Kruskal-Wallis (1645) and low correlation-cutoff of 0.25 left 1065 predictors in this study.
- Then examined the performance of LASSO, Elastic Net, LDA and Classification Tree models, using genetic profiles and patients health status as predictors.
- Our application of a pruned Classification Tree did not net comparable results on this data. LDA was consistently the penultimate.
- The best model was the Elastic Net with $\alpha = 0.3$, having accuracy of 0.983 and F_1 of 0.982.
- The Elastic Nets with different parameters have comparable performance (including LASSO).

University of Victoria