Expanding Vitis Vinifera flavonoid network using Feature Selection Methods

Aliaksandr Siarohin, Mat. number 180402
aliaksandr.siarohin@studenti.unitn.it

I. Introduction

In this project we had been given expression data of vitus vinifera genes, and part of flavonoid pathway. We had been asked to expand this pathway. One way of doing it is to apply feature selection methods. In this project I tried 3 methods: lasso [1], randomized lasso [2] and recursive feature elimination [3] with random forest [4].

II. Preprocesing

Expression data contains a lot of NaN values, and in order to apply feature selection methods we first need to deal with this NaNs. I decided to remove all the experiments with amount of NaNs > 5000. And then replace NaNs that left with mean, computed across all genes in given experiment (e.g column mean). The number of removed/replaced can be found in Table I.

Table I
Preprocesing statistic. Show how many genes and experiments I remove after preprocessing, and how many NaN values we replace with column mean.

Removed Experiments	Removed Genes	Replaced With Mean NaN
863	0	613890

III. Method Description

Let G - be set of genes, E - set of experiments, M - gene expression matrix. For now let‘s assume that we want to predict genes expression of gene g, based on all the other genes. We can construct following regression problem $X = M^T_{E,G/\{g\}}$, $Y = M^T_{E,g}$, so the regression problem is to predict Y given X. Then using feature selection methods we want to obtain genes that explain (predict) g, and ranking of this genes.

A. Lasso

Lasso method as described in [1] has regularization parameter α which controls amount of regularization. I start from 0.1 and try to decrease this parameter until method select 100 genes (The lower α increase computational time, so 100 genes is a trade-of between efficiency and accuracy). The ranking of genes can be obtaining using absolute value of weight assigned to every selected gene.

B. Randomized lasso

Randomized lasso as described in [2] run lasso on random subsets of training set (I use 30 runs, and 75 % of training set for every run). So I use the same setting for estimating alpha (start from 0.1 and decrease until method select 100 genes). The ranking can be obtained from the number of times that method select particular gene.

C. Recursive feature elimination with Random forest

Recursive feature elimination as described in [3] fit regression method on all dataset, than discard feature with the most lowest rating, fit regression on reduced dataset, discard feature with lowest rating and so on, until one feature is left. Random forest on the other hand can produce features ranking, based on the number of times the feature selected when random forest was built. The problem of this method is that for all the genes it take long time to run. So I decided to take genes selected by Lasso method. So this method just assign new ranks. I use random forest with 100 trees in it. And assigned rank using formula:

$$\frac{1}{n\_features - step\_at\_which\_feature\_was\_discarded}$$

I ran this 3 methods for every target gene in the flavonoid pathway.

D. Method Comparison

To compare methods I select 100 top rated genes and use 2 metrics average: number of genes from pathway selected by method and average DCG[1](gene from pathway get relevance equal to 1, and not from pathway get relevance equal to zero). The result is in Table II. From this table we can see that all methods show more or less the same performance.

IV. Final Result

To obtaining final result I use rank aggregation described in [5]. I aggregate list for all methods and for all genes, and select best 25. Highest rated get highest rank. The results can be found in Table III and in Table IV.

References

[1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[1] https://en.wikipedia.org/wiki/Discounted_cumulative_gain

Table II
FEATURES SELECTION METHOD COMPARISON.

Method	Average number of pathway genes	Average DCG
Lasso	3,23	1,10
RFE	3,23	1,03
RandomizedLasso	2,98	0,87

Table III
TOP10 FINAL GENES SELECTED BY AGGREGATING RESULTS OF ALL METHODS.

Rank	Gene	Is in pathway
24	VIT_14s0068g00920	True
23	VIT_06s0004g02620	False
22	VIT_18s0001g09400	False
21	VIT_04s0023g03370	True
20	VIT_03s0088g00260	False
19	VIT_15s0046g00170	True
18	VIT_18s0001g14310	True
17	VIT_00s0361g00040	True
16	VIT_06s0009g03040 VIT_06s...	True
15	VIT_01s0010g03510	True

[2] N. Meinshausen and P. Bühlmann, “Stability selection,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 4, pp. 417–473, 2010.

[3] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[4] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] V. Pihur, S. Datta, and S. Datta, “Weighted rank aggregation of cluster validation measures: a monte carlo cross-entropy approach,” *Bioinformatics*, vol. 23, no. 13, pp. 1607–1615, 2007.

Table IV
TOP10 FROM FINAL GENES, THAT ARE NOT IN PATHWAY, SELECTED BY AGGREGATING RESULTS OF ALL METHODS.

Rank	Gene
23	VIT_06s0004g02620
22	VIT_18s0001g09400
20	VIT_03s0088g00260
8	VIT_14s0060g00900
6	VIT_13s0064g01480
5	VIT_02s0012g01570
4	VIT_16s0100g00270
2	VIT_12s0028g01150
1	VIT_03s0091g01290
0	VIT_07s0005g06460