On the performance of variable selection and classification via ranked based classifier

Showaib Rahaman Sarker
The University of Texas, USA

In high-dimensional gene expression data analysis, the accuracy and reliability of cancer classification and selection of important genes play a very crucial role. To identify these important genes and predict future outcomes (tumor vs. non-tumor), various methods have been proposed in the literature. But only few of them take into account correlation patterns and grouping effects among the genes. In this article, we propose a rank-based modification of the popular penalized logistic regression procedure based on a combination of l1 and l2 penalties capable of handling possible correlation among genes in different groups. While the l1 penalty maintains sparsity, the l2 penalty induces smoothness based on the information from the Laplacian matrix, which represents the correlation pattern among genes. We combined logistic regression with the BH-FDR (Benjamini and Hochberg false discovery rate) screening procedure and a newly developed rank-based selection method to come up with an optimal model retaining the important genes. Through simulation studies and real-world application to high-dimensional colon cancer gene expression data, we demonstrated that the proposed rank-based method outperforms such currently popular methods as lasso, adaptive lasso and elastic net when applied both to gene selection and classification.

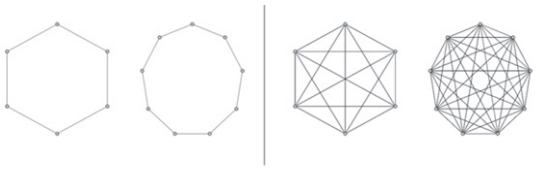

Figure 1. The ring network (left) and F.con network (right) are shown for the case there are two genes consisting of 6 and 9 CpG sites, respectively.

Table 9. List of top 5 ranked genes across rank-based, lasso, adaptive and elastic net. An extra asterix (*) sign is put next to a gene each time the gene is selected by one of four methods.

EST Name	Gene ID	Gene Description	Selection Probability
Rank-Based			
***Hsa.36689	Z50753	H.sapiens mRNA for GCAP-II/uroguanylin precursor	1.00
***Hsa.692.2	M76378	Human cysteine-rich protein (CRP) gene, exons 5 and 6	0.99
**Hsa.37937	R87126	Myosin heavy chain,nonmuscle(Gallus gallus)	0.97
****Hsa.1660	H55916	Peptidyl-prolyl cis-trans isomerase, mitrochondrial precursor(human)	0.91
Hsa.1832	R44887	nedd5 protein (Mus musculus)	0.90
Lasso			
***Hsa.36689	Z50753	H.sapiens mRNA for GCAP-II/uroguanylin precursor	0.87
Hsa.692.2	M76378	Human cysteine-rich protein (CRP) gene, exons 5 and 6	0.82
*****Hsa.1660	H55916	Peptidyl-prolyl cis-trans isomerase, mitrochondrial precursor(human)	0.66
Hsa.6814	H08393	Collagen alpha 2(XI) chain(Homo sapiens)	0.52
Hsa.8147	M63391	Human desmin gene, complete cds	0.50
Adaptive Lasso			
Hsa.1454	M82919	H. gamma amino butyric acid(GABAA)receptor beta3 subunit mRNA,cds	0.83
Hsa.6814	H08393	Collagen alpha 2(XI) chain(Homo sapiens)	0.77
****Hsa.1660	H55916	Peptidyl-prolyl cis-trans isomerase, mitrochondrial precursor(human)	0.77
Hsa.14069	T67077	Sodium/Potasssium-transporting atpase gamma chain(Ovis aries)	0.69
Hsa.2456	U25138	Human MaxiK potassium channel beta subunit mRNA, complete cds	0.55
Elastic Net			
***Hsa.36689	Z50753	H.sapiens mRNA for GCAP-II/uroguanylin precursor	0.98
**Hsa.37937	R87126	Myosin heavy chain,nonmuscle(Gallus gallus)	0.94
***Hsa.692.2	M76378	Human cysteine-rich protein (CRP) gene, exons 5 and 6	0.94
Hsa.8147	M63391	Human desmin gene, complete cds	0.91
****Hsa.1660	H55916	Peptidyl-prolyl cis-trans isomerase, mitrochondrial precursor(human)	0.84

Biography

Showaib Rahman Sarker is pursuing his master's degree in Statistics at The University of Texas at El Paso. He has expertise in Statistical Machine learning application in High-Dimensional Gene Expression Data. Currently, he is doing his research in High throughput cancer gene expression data. His main goal is to find out the important genes which are responsible for cancer and classify (tumor vs non-tumor) accurately. He is passionate to apply statistical approach and machine learning approach in cancer research.

msarker@miners.utep.edu