Gene Expression Analysis - Assignment 3

Franco Chiesa Docampo and Luis Felipe Aycardi

Bioinformatics - LGBIO2010

1 Introduction

In this assignment we will analyze gene expression data and gradually select the most informative markers for a diagnosis task using Support-Vector Machines. The data is composed by 54613 probeset expression values from microarrays of 103 different patients, suffering from Glioblastoma. We search in particular for expression markers for predicting whether a patient suffers from Glioblastoma or not. gplots, mRMRe and LiblineaR libraries to data classification and plotting are used in R.

In the present work we pre-process, normalize and filter out certain features according to statistical criteria.

2 Non-specific filtering

A first step to perform some sort of filtering, is to rank the various probesets by increasing variance on all samples. From them, we will Keep only 25 percent of probesets with the larger variances, meaning the ones that change the most along the samples. The number of probesets kept after this non-specific filtering is 13654 and the minimal variance considered for a probeset to be kept is 0.5085179.

Another configuration, not as straightforward to select this 25 percent (in which this exact number would not be compute) is to rank the probset according to its variances in increasing order and find the minimum and maximum variances. Then, find the 75 percent of the variances and match the gene that corresponds to that one. From it, finally select all genes with a higher variance and keep them.

Following this, the names and variances of the top 5 probesets with the largest variances are presented in table 1.

Table 1: Top 5 probesets with the largest variances.

Rank	Gene	Variance
1	X 50897	40.83417
2	X 5690	33.81839
3	X 7692	15.28151
4	X 1355	12.01631
5	X 33565	11.92148

The plot of the ranked variances from each probeset is as follows in Figure 1.

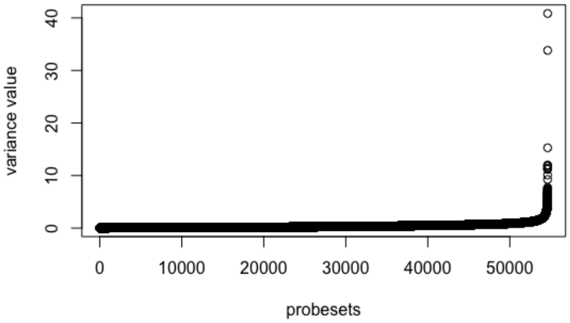

Figure 1: Probeset variances by increasing ordered.

We believe that this plot behaviour makes sense because (as the variance values are ranked by increasing ordered) the variance numbers increase gradually as the right limit is approached. The main difference between the plot displayed in the assignment and the one that we produced is that ours is more ”flat” until the last 600 genes (probesets) appear. This is where the variance plot ”takes off” and starts to show bigger values. The characteristic ”S” shape of the plot showed as an example is also present in our plot, the thing is that the last 600 probesets are so big in variance value compared to the rest that the ”S” shape almost disappears to the human eye. Functions useful for the making of this plot is algorithm ”order”, which offers a vector that contains the original vector position of the variance values so that we can easily create a new ranked vector.

For the rest of this assignment, we will only consider the dataset restricted to those features that have not been filtered out at this step. But to make sure that each gene (probeset) has roughly the same expression range across all samples, a normalization procedure is needed. It consists in centering a feature on its mean value and dividing it by its standard deviation. With these data, we will develop the rest of the filtering and procedures in the assignment.

3 Differential expressed probesets

To perform a differential expressed analysis, the next step is to rank the various probesets using a t-test by increasing p-values to distinguish between the two conditions under study (Glioblastoma or not). The ranking produced is depicted in Figure 2.

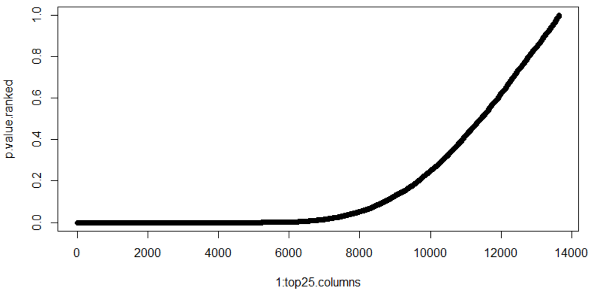

Figure 2: Probeset p-values by increasing ordered.

According to the t-test performed, a total of 7954 features are statistically significant, consider a 5% significance level and no specific correction at this point.

However, by looking at the probability of type I error of a statistical test, we could conclude that the mean expression values among the 2 classes are significantly different for a given gene while they are not. As this leas to a false selection of feature with probability α some corrections can be done.

A Bonferroni correction of the original t-test will divide the critical value (α= .05) by the number of tests nt performed, which is usually very conservative. For this probeset, 4125 features are considered significant with this correction.

The top 10 most discriminating features and the associated corrected p-values in increasing order are presented in table 4.

Table 2: Top 10 most discriminating features with Bonferroni correction.

Rank	Gene	corrected p-value
1	X 22349	2.132537e-31
2	X 27963	5.019396e-31
3	X 27904	1.044625e-30
4	X 19483	1.654585e-29
5	X 29037	2.364043e-29
6	X 37364	2.705285e-29
7	X 37067	3.417921e-29
8	X 27337	3.673879e-29
9	X 21839	5.494453e-29
10	X 12761	1.425342e-28

Another approach could be to consider more features in the correction, which leads to a False Discovery Rate (FDR) correction of the original t-test, which normally leads to select all features. For this probeset, 7439 features are considered significant with this correction.

The top 10 most discriminating features and the associated corrected p-values in increasing order are presented in table 3.

As it is presented in tables 4 and 3, those corrections do not change the relative ranking of features, just the selection threshold. In other words,

Table 3: Top 10 most discriminating features with FDR correction.

Rank	Gene	corrected p-value
1	X 22349	2.132537e-31
2	X 27963	2.509698e-31
3	X 27904	3.482082e-31
4	X 19483	4.136463e-30
5	X 29037	4.508808e-30
6	X 37364	4.508808e-30
7	X 37067	4.592349e-30
8	X 27337	4.592349e-30
9	X 21839	6.104948e-30
10	X 12761	1.425342e-29

a FDR correction is equivalent to Bonferonni correction whenever a single feature is selected. Therefore, then selected genes are the same although the corrected p-value relatively change.

4 Heatmaps and data visualization

A heatmap is produced by agglomerative clustering (see phylogeny) respectively on rows (samples) or columns (genes). A heatmap of all samples along the 50 most differential expressed features is presented in Figure 5.

As it is evident in the heatmap that some clusters are present. We could debate around the precise number of clusters that one can identify, but in the end the fact is that there are different groups. Such a configuration is expected since we are considering the features that showed statistical differences among classes (Glioblastoma or not). If they are the features that best express a difference in the label, the level of expression will be depicted in such a map.

A 2-D plot of all samples along the 2 most significant features according to the differential expression analysis is presented in Figure 3.

To label each sample with its respective condition, specific color coding was used. In red are plotted the samples with Glioblastoma and in blue, the ones with No tumor.

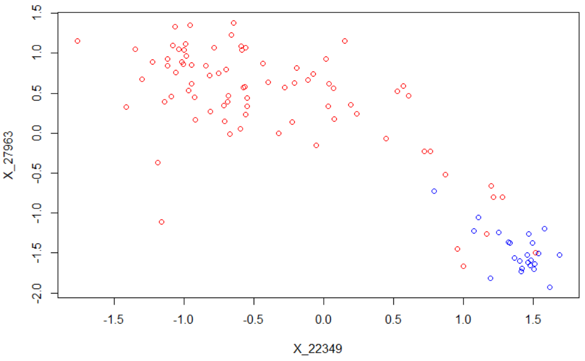

Figure 3: 2 most significant features.

From the previous plot we can observe that the 2 conditions can be, not perfectly, but fairly well distinguished. Even though there are some sample overlapping, the core of each group can be easily spotted. This is consistent with the previous heatmap as a cluster can be generated for each of the classes. It is therefore, a reduced representation of what was described in previous analysis.

On the other hand, a 2-D plot of all samples along the 2 least significant features among those kept after the non-specific filtering is presented in Figure 4. The same color coding of the previous figure is followed.

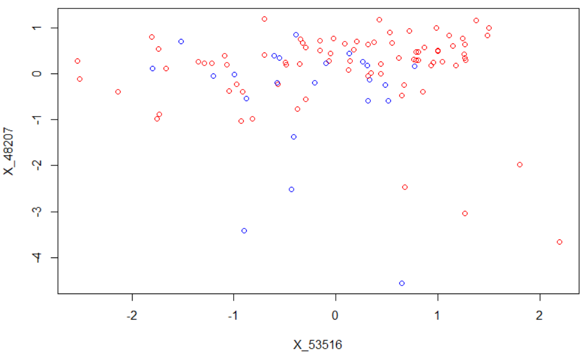

Figure 4: 2 least significant features.

In this case, and as expected, the two conditions can not be well distinguished. In fact, the distribution of them is very similar and no clusters can be done. In contrast to what is expressed in the heatmap and with the 2 most significant features, genes do not provide substantial information to make a classification.

5 Predictive Models

We proceed to split the available samples into a 80 percent training and 20 percent test. Which gives us a training set of 82 samples vs a test set of 21 samples. It is important to stick to this training/test partition to avoid over-fitting of the model and not having a biased performance result. With more training data, the parameter estimates will have a smaller variance. If the size of the testing data is bigger, the performance statistic will have smaller variance. There is a compromise relationship between training and testing data.

We proceed to fit a linear SVM on the training using all probesets obtained after non-specific filtering performed in section 2. A table with the 10 most important features according to the absolute weight values of such a linear model is presented in table 4.

Table 4: Top 10 features according to SVM weight.

Gene	Weight	p-value rank
X 16482	0.33680499	1989
X 34099	0.08517068	1987
X 4862	0.07350973	3455
X 44516	0.07014985	2838
X 4360	0.05703408	2894
X 53080	0.05256469	3087
X 16609	0.05109622	2640
X 6322	0.05052760	9830
X 4495	0.04630046	4282
X 756	0.04601871	5386

The absolute size of the weight coefficient relative to the other ones indicates the importance of such feature for performing the plane separation in order to do the classification task. The 10 most important features of the SVM model do not match the most differential expressed probesets (according to the p-value ranking). 7 of the 10 weights represented in this table are inside the Bonferroni correction, which (as stated previously) is a conservative selection. Therefore we can conclude that the genes that have a high SVM weight are likely to be statistically significant according to their p-value.

The confusion matrix produced by the algorithm is presented in table 5.

Table 5: Confusion matrix.

	predicted T	predicted F
actual T	12	1
actual F	1	7

The classification accuracy of the SVM is of 0.8990385. Which makes sense because as we could observe in Figure 3 and 4, there are dots (pacients/observations) which correspond to a certain class that are among other dots that are labelled differently. Resulting in some degree of error in the separation of the data by the hyperplane.

6 Feature selection

If a gene has random expressions or is uniformly distributed in different classes, its mutual information with these classes is zero. If a gene is strongly differentially expressed for different classes, it should have large mutual information. Thus, we use mutual information as a measure of relevance of genes.

Now we will compute the mutual information between each feature and the class label samples. Table 6 present (i) the ranking of each feature (from 1 to 10), (ii) the feature name and (iii) the MI value associated.

Now we will select the 10 most relevant features according to the mRMR algorithm and report a table containing the rank of each feature (here the rank denotes the iteration of the mRMR algorithm where such a feature is selected), the feature name and the associated mRMR scores (table 7)

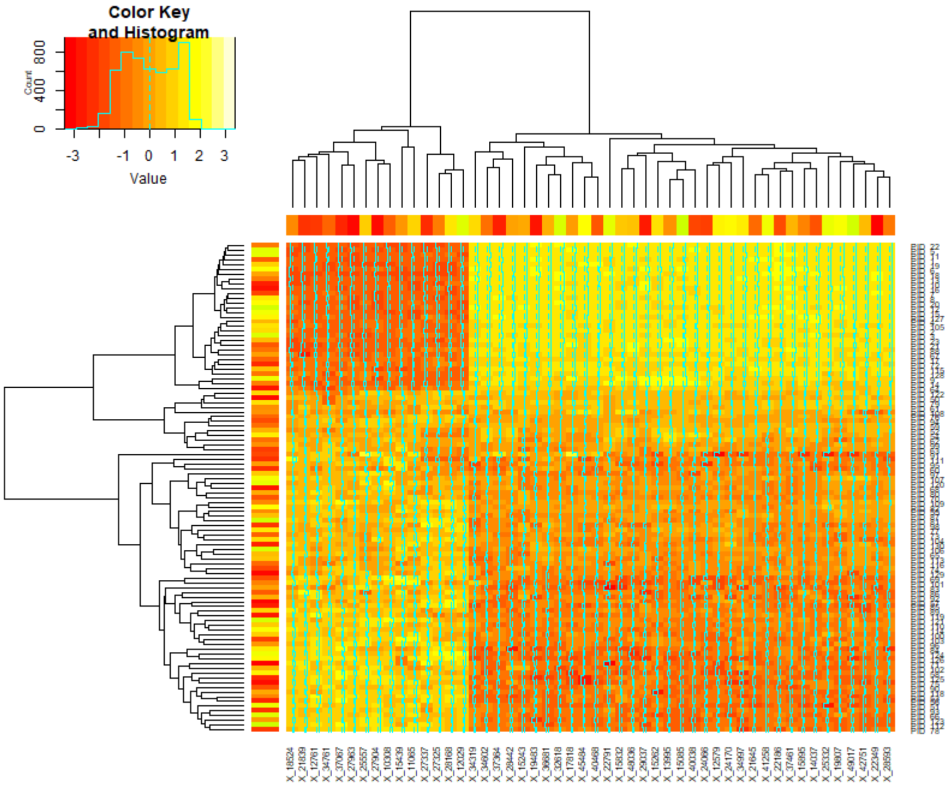

Figure 5: Heatmap of all samples along the 50 most differential expressed features.

Table 6: Top 10 features according to Mutual Information.

Rank	Gene	MI
1	X 16482	0.6895056
2	X 13610	0.6155210
3	X 31845	0.6001876
4	X 10741	0.5890955
5	X 24315	0.5674374
6	X 10740	0.5480804
7	X 3205	0.5098927
8	X 23990	0.5012261
9	X 42398	0.4992005
10	X 27963	0.4972281

Table 7: Top 10 features according to the mRMR algorithm

mRMR Gene	Rank	mRMR score
X 12593	6706	0.09325788
X 36359	6998	0.09280914
X 24315	13069	0.10627251
X 16609	9283	0.09877292
X 22100	1170	0.09447220
X 13610	13536	0.09853090
X 31845	9535	0.13321722
X 47539	4262	0.08284041
X 44516	10346	0.07592211
X 16482	12880	0.68950565

Minimum redundancy feature selection is used in a method to identify characteristics of genes and find their relevance. mRMR is calculated by pairing relevant features to the 2 classification variables displayed in the ”label” column.

7 References

Extracts of our development are based on information from:

- https://www.rdocumentation.org/packages/LiblineaR/versions/2.10-8/topics/LiblineaR

- http://ranger.uta.edu/chqding/papers/gene_select.pdf