# Carrak - User Manual

Camille Maumet

July 18, 2008

# Contents

# Chapter 1

# Introduction

This software provides an interface to train classifiers and to estimate the predictive error rate of classifiers using external one-layer and two-layer cross-validation . These two cross-validation techniques have been presented respectively in [Ambroise and McLachlan, 2002] and [Stone, 1974], [Wood et al., 2007], [Zhu et al., 2008]. One-layer external cross-validation can be used to determine a nearly unbiased estimate of the error rate in the context of feature selection. The number of features to be selected is specified and feature selection is performed on a set of training folds, with the error rate estimated across the test folds. This procedure is repeated for each number of features to be selected. As an output of this one-layer cross-validation, the user gets a cross-validated error rate per size of subset. However, if the user wants to know the smallest estimated error rate over all the subsets considered, then a second layer of cross-validation is required to estimate the effect of this choice.

This document describes how to install this software and make the best use of its functionality.

# Chapter 2

# Installation and dependencies

## 2.1 Dependencies

This software depends upon the following other software that has to be installed already, please check that all of these are installed before starting Carrak:

- R with a version later or equal to 2.6.2, currently not available for version 2.7.1
- RMagpie R package, available at http://www.maths.uq.edu.au/bioinformatics/rmagpie.html

## 2.2 Installation

Please follow the instructions corresponding to your platform.

For Windows, download the zip sources from ... and unzip the folder (right click and choose for example, `extract here`). Then, simply double click on `carrak.exe`.

# Chapter 3

# Home page

The home page, as depicted in figure 3.1 gives access to the main functionalities of our dataset. From here you can:

- Access your datasets and upload new datasets.
- Run one-layer and two-layer cross-validation with various types of classifiers and feature selection, estimate the predictive error rate on a dataset.
- Access your previous experiments and display them.
- Create a classifier which can be used to classify new data.

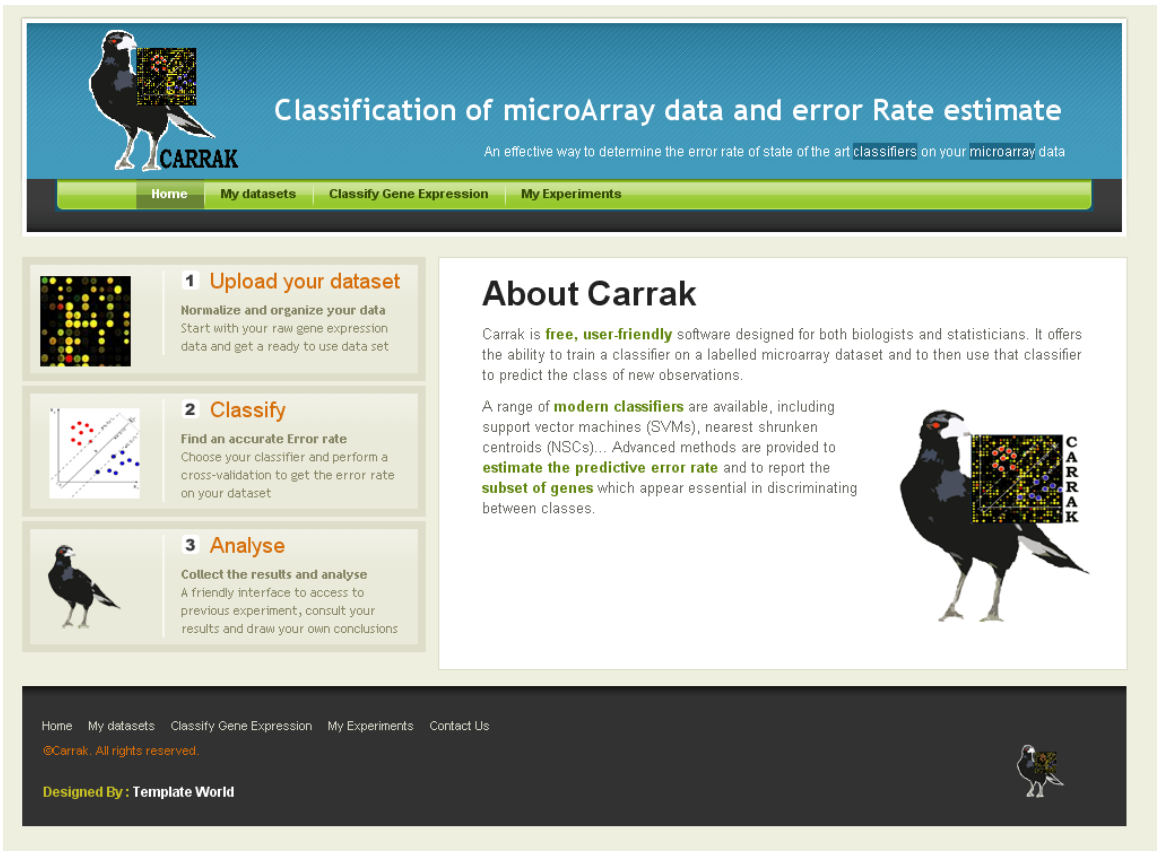



Figure 3.1: Home page

# Chapter 4

# Data sets

## 4.1 General Presentation

By clicking one the `My Datasets` entry of the menu you get a list of the data sets currently available, as displayed in figure 4.1.

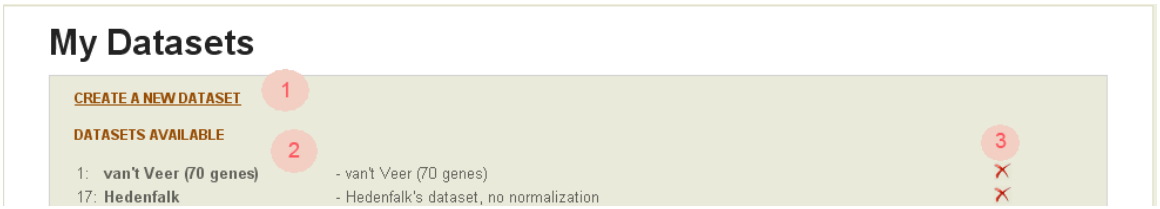



Figure 4.1: List of your data sets

The list display the identifier of each data set, its name and description. From here you can:

1. Upload a new data set by clicking on 'CREATE A NEW DATASET'
2. Delete a data set by clicking on the red cross

These actions are described in the next sections.

## 4.2 Upload a new data set

To upload a new data set you must fill the form displayed in figure 4.2 and give the following pieces of information:

- Choose a name for your data set
- Write a small description
- Specify whether or not the names of the samples are available in the files. Be careful if the names of the samples are given, they must appear in both files. If they are not they will be automatically generated as 'V1', 'V2'...
- Specify whether or not the names of the genes are provided in your gene expression file. If they are not, they will be automatically generated as 'gene1', 'gene2'...
- Give the path and name of the file containing the levels of gene expression of each gene for each sample.
- Specify whether or not you want your data to be normalized (as a normal distribution with mean zero and standard deviation one.)

- Usually, each row in your gene expression file corresponds to a gene and each column to a sample. If it's not the case you must answer 'No' to the question: 'Does a column correspond to a sample and a row to the expression levels for a given gene ?'
- Give the path and name of the file containing the class label of each sample.
- Specify whether the class labels are order on a row or a column in your class label file.
- When you are ready, click on 'Create'

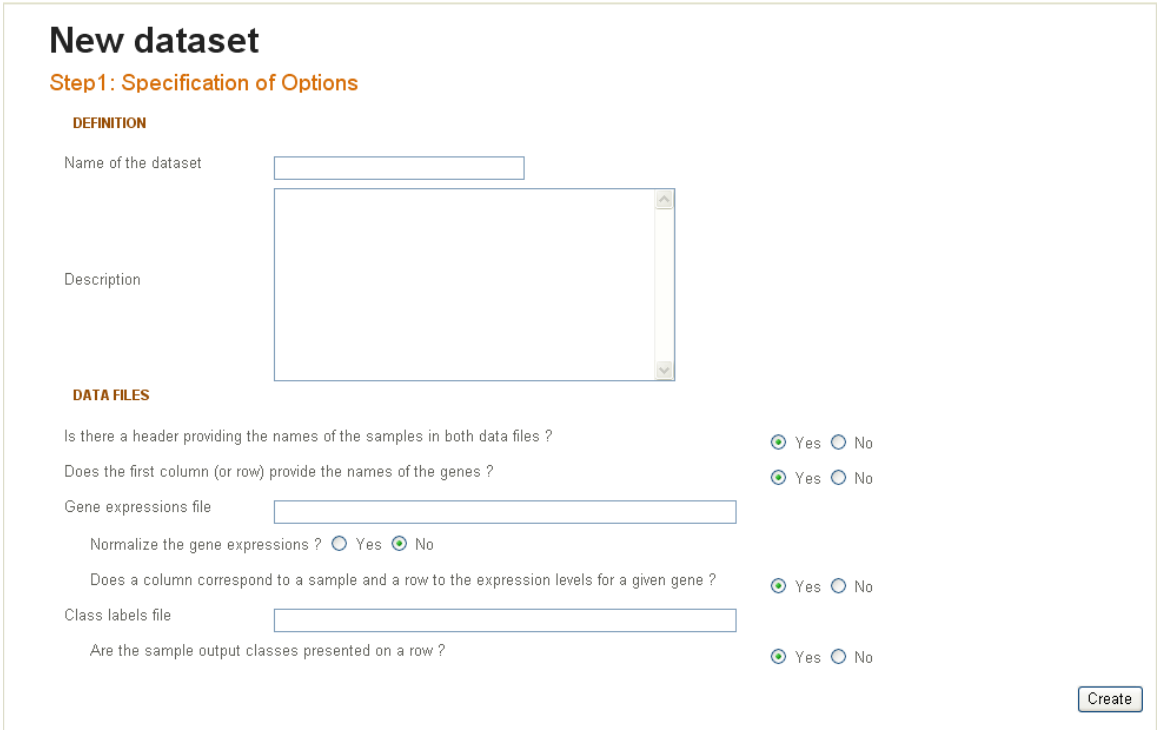



Figure 4.2: Upload a new data set

## 4.3 Delete an existing data set

To delete an existing data set, click on the red cross corresponding to the one you want to delete from the list of data sets. As displayed in figure 4.3, a message wait for your confirmation before deleting the selected data set. Click on `ok` to confirm the deletion or on `Cancel` if you want to go back the the list of data sets without deleting the selecting data set.

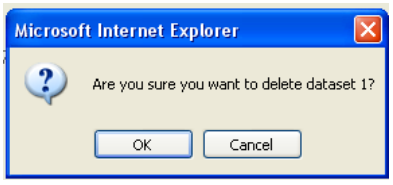



Figure 4.3: Deletion of an existing data set

# Chapter 5

# Classification

By clicking one the `Classify Gene Expression` entry of the menu you get form as displayed in figure 5. This form will allow you to perform a one-layer and a two-layer cross-validation by a simple click.

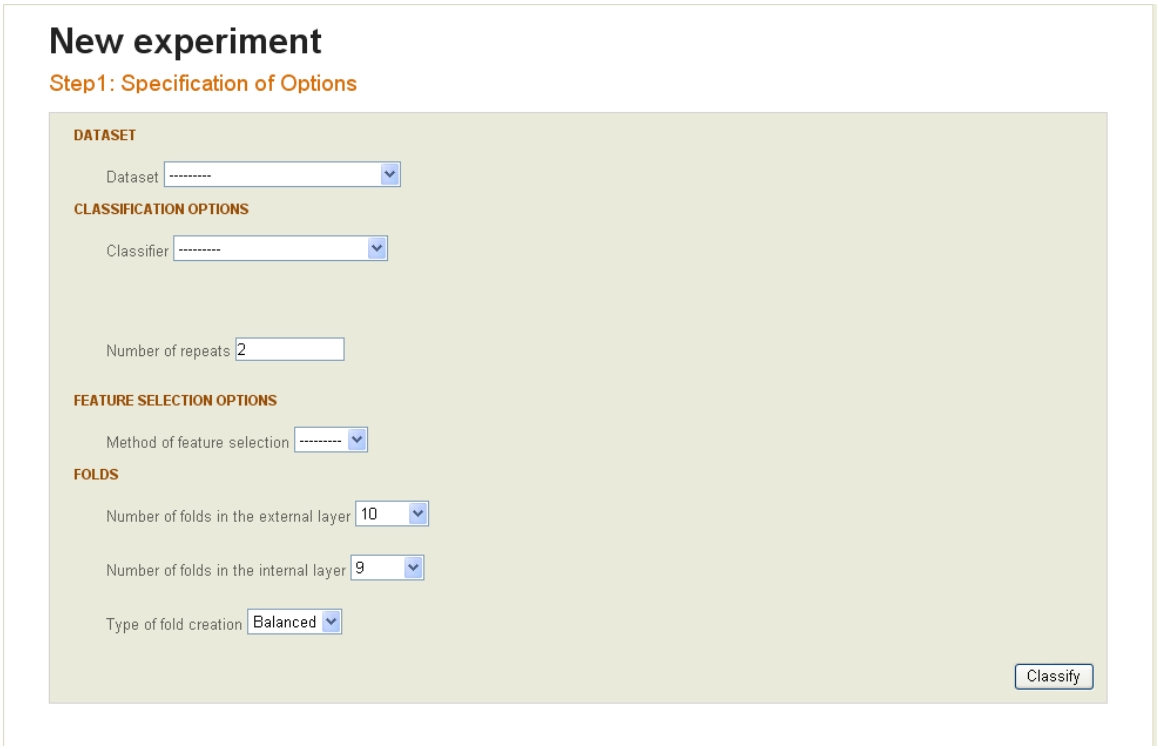



Figure 5.1: List of your data sets

Let's have a closer look at what pieces of information you need to fill this form:

- First, you have to choose on which data set the cross-validations will be applied. If no data set appears in the selection box that means that you need to upload a data set before. To do so, please have a look at section 4.2.
- Second, you can choose between two classifiers, namely the Support Vector Machine with the Recursive Feature Elimination proposed in [Guyon et al., 2002] or the Nearest Shrunken Centroid presnted in [Tibshirani et al., 2002].
- If you have chosen the Support Vector Machine, you need to choose the kernel. In a first approach, the 'Linear' is probably the most appropriate. Also, it's the fastest kernel to be processed.

- You also need to choose the number of repeats. Each cross-validation (one-layer or two-layer) will be performed several times according to the number of repeats given here. The results are then averaged over the repeats. It is believed, [Burman, 1989], that this method can improve the accuracy of the estimator of the error rate.
- You then have to choose the sizes of subsets (for RFE) or the thresholds (for NSC) that will be tried. By default subsets from size one to one to the number total of features by powers of two are tried for the RFE or 30 thresholds selected by the pamr package for the NSC. You can also select you own sizes or thresholds.
- Then, choose the number of folds that you want in the outer layer of two-layer cross-validation and for one-layer cross-validation. The default value is 10.
- Choose the number of folds that you want in the inner layer of two-layer cross-validation. The default value is 9.
- The last step is to specify, the kind of division in fold that you want. Two options are possible: a simple way or balanced folds as presented in [] CITATION MISSING

When you are ready click on classify and wait until your experiment is finished. The results are then displayed as presented in section 6.2.

# Chapter 6

# Experiments

## 6.1 General presentation

By clicking one the `My Experiments` entry of the menu you get a list of the experiments currently available, as displayed in figure 6.1.

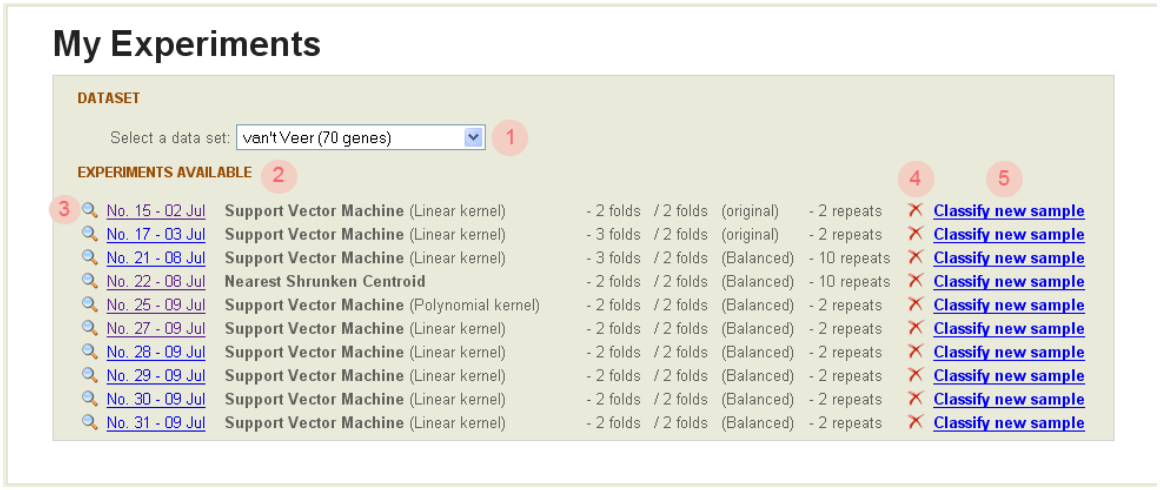



Figure 6.1: Home page

From this view you can:

1. Access to the list of experiments made on a given data set by selecting a data set
2. Look at the list of all the experiments available for the selected data set
3. Consult an experiment by clicking on its identifier or on the magnifying glass
4. Delete an experiment by clicking on the red cross
5. Classify new samples by clicking on the corresponding link

The three last actions are described in the next sections.

## 6.2 Consult an experiment

The window displaying the results of an experiment is divided into two parts:

- The 'Options of the Experiment' section reminds the user of the options used to compute the classification (Classifier name, dataset...)
- The 'Experimental results' describe the results obtained by two-layer and one-layer cross-validation.

Both parts are described in further details in section 6.2.1 and 6.2.2. Figure 6.2 presents an overview of the interface.

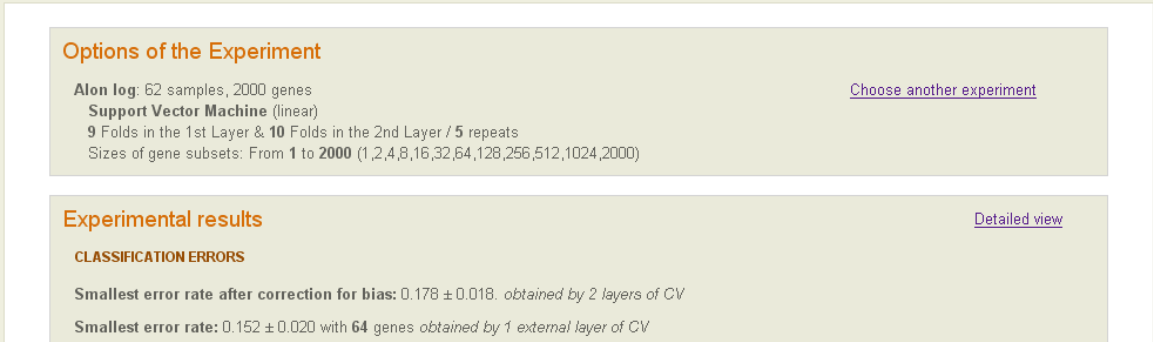



Figure 6.2: Results of an experiment

### 6.2.1 Options

The options are displayed in a frame at the top of the window. It includes all the information depending on the experiment currently displayed:

- Dataset name and short description (number of samples and genes)
- Classifier name
- Number of folds in the outer (2nd layer) and inner (1st layer) layers
- Number of repeats of cross-validation
- Different sizes of subsets or thresholds tried during the gene selection

The 'Choose another experiment' link brings the user back to the window displaying the list of the experiments where he can select another experiment.

### 6.2.2 Experimental Results

This part is divided in a 'Summary view' and a 'Detailed view' that are accessible from the links 'Summary view' and 'Detailed view' located at the top of the results part (cf. Figure 6.3). By default the 'Summary view' is displayed. Section 6.2.2 will describe the summary view and section 6.2.2 present the detailed view.

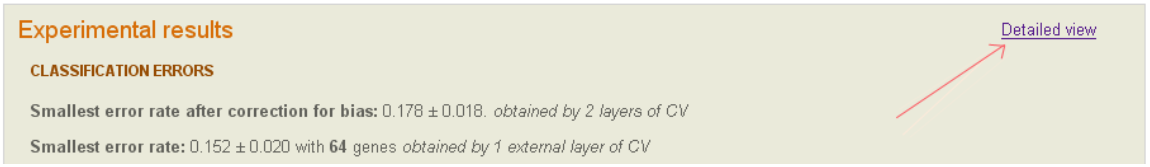



Figure 6.3: Experimental results: Choice between summary or detailed view

### Summary view

This view is divided in two parts, at the top, under the title 'Classification errors', are the information concerning the error rate and at the bottom, under the title 'Genes selected', the one regarding the genes selected by feature selection during the classification process.

**Classification errors** We will first have a look at the 'Error rate' part. Figure 6.4 presents the five interesting points displayed in this view:

1. The biased-corrected error rate obtained by two layers of cross-validation
2. The best number of genes and corresponding error rate obtained by one layer of cross-validation
3. A graph showing the one-layer cross-validated error rate versus the number of genes in the subset (usually displayed as logarithm of the number of genes)
4. A table summarizing the cross-validated error rate versus the number of genes in the subset and optionally the error rate per class.
5. These check-boxes allow the user to display or hide the class error rates. Figure 6.5 give an example of the possible displays.

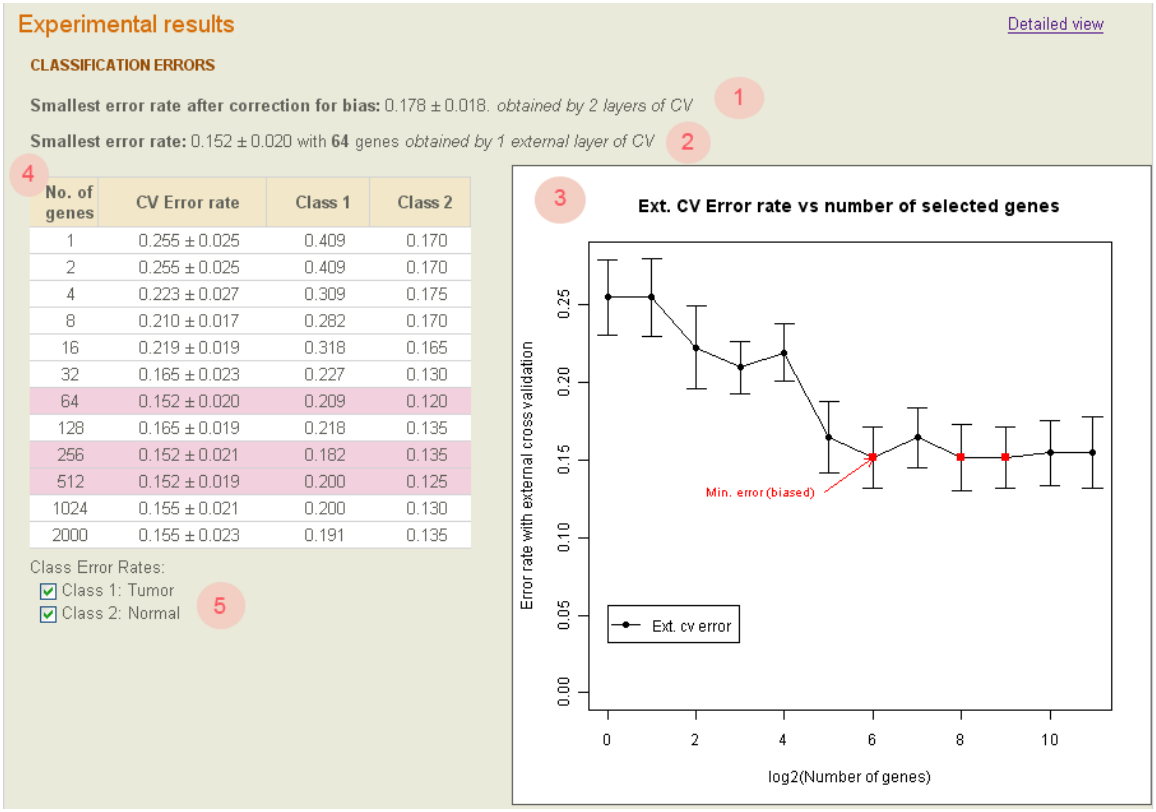
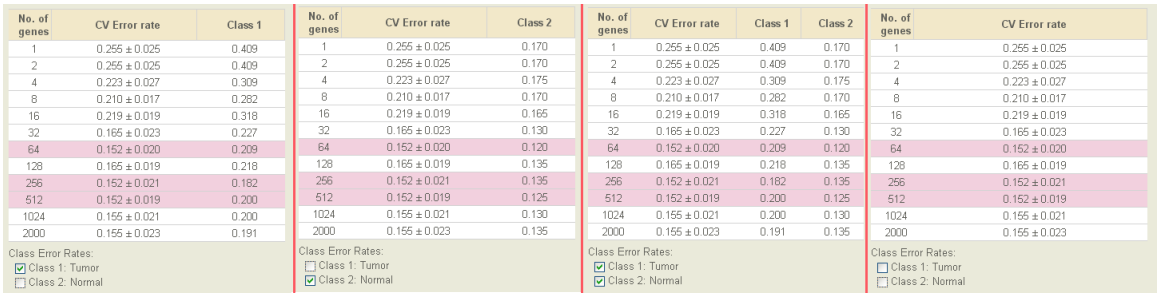



Figure 6.4: Summary view: Classification errors

**Genes Selected** The second part of the view, displays the gene selected during the classification and their relevance to the given outcome. Figure 6.6 points out the two different visualization tools to display the frequency of selection of the genes among the subsets and repeats:

1. A table where the genes selected are ordered by frequency for each size of subset or threshold
2. A microarray-like image

Diverse tools enable the user to adapt the display to his need.



| No. of genes | CV Error rate | Class 1 |
|---|---|---|
| 1 | 0.255 ± 0.025 | 0.409 |
| 2 | 0.255 ± 0.025 | 0.409 |
| 4 | 0.223 ± 0.027 | 0.309 |
| 8 | 0.210 ± 0.017 | 0.282 |
| 16 | 0.219 ± 0.019 | 0.318 |
| 32 | 0.165 ± 0.023 | 0.227 |
| 64 | 0.152 ± 0.020 | 0.209 |
| 128 | 0.165 ± 0.019 | 0.218 |
| 256 | 0.152 ± 0.021 | 0.182 |
| 512 | 0.152 ± 0.019 | 0.200 |
| 1024 | 0.155 ± 0.021 | 0.200 |
| 2000 | 0.155 ± 0.023 | 0.191 |

Class Error Rates:
Class 1: Tumor
Class 2: Normal

| No. of genes | CV Error rate | Class 2 |
|---|---|---|
| 1 | 0.255 ± 0.025 | 0.170 |
| 2 | 0.255 ± 0.025 | 0.170 |
| 4 | 0.223 ± 0.027 | 0.175 |
| 8 | 0.210 ± 0.017 | 0.170 |
| 16 | 0.219 ± 0.019 | 0.165 |
| 32 | 0.165 ± 0.023 | 0.130 |
| 64 | 0.152 ± 0.020 | 0.120 |
| 128 | 0.165 ± 0.019 | 0.135 |
| 256 | 0.152 ± 0.021 | 0.135 |
| 512 | 0.152 ± 0.019 | 0.125 |
| 1024 | 0.155 ± 0.021 | 0.130 |
| 2000 | 0.155 ± 0.023 | 0.135 |

Class Error Rates:
Class 1: Tumor
Class 2: Normal

| No. of genes | CV Error rate | Class 1 | Class 2 |
|---|---|---|---|
| 1 | 0.255 ± 0.025 | 0.409 | 0.170 |
| 2 | 0.255 ± 0.025 | 0.409 | 0.170 |
| 4 | 0.223 ± 0.027 | 0.309 | 0.175 |
| 8 | 0.210 ± 0.017 | 0.282 | 0.170 |
| 16 | 0.219 ± 0.019 | 0.318 | 0.165 |
| 32 | 0.165 ± 0.023 | 0.227 | 0.130 |
| 64 | 0.152 ± 0.020 | 0.209 | 0.120 |
| 128 | 0.165 ± 0.019 | 0.218 | 0.135 |
| 256 | 0.152 ± 0.021 | 0.182 | 0.135 |
| 512 | 0.152 ± 0.019 | 0.200 | 0.125 |
| 1024 | 0.155 ± 0.021 | 0.200 | 0.130 |
| 2000 | 0.155 ± 0.023 | 0.191 | 0.135 |

Class Error Rates:
Class 1: Tumor
Class 2: Normal

| No. of genes | CV Error rate |
|---|---|
| 1 | 0.255 ± 0.025 |
| 2 | 0.255 ± 0.025 |
| 4 | 0.223 ± 0.027 |
| 8 | 0.210 ± 0.017 |
| 16 | 0.219 ± 0.019 |
| 32 | 0.165 ± 0.023 |
| 64 | 0.152 ± 0.020 |
| 128 | 0.165 ± 0.019 |
| 256 | 0.152 ± 0.021 |
| 512 | 0.152 ± 0.019 |
| 1024 | 0.155 ± 0.021 |
| 2000 | 0.155 ± 0.023 |

Class Error Rates:
Class 1: Tumor
Class 2: Normal

Figure 6.5: Classification errors: Class error rates

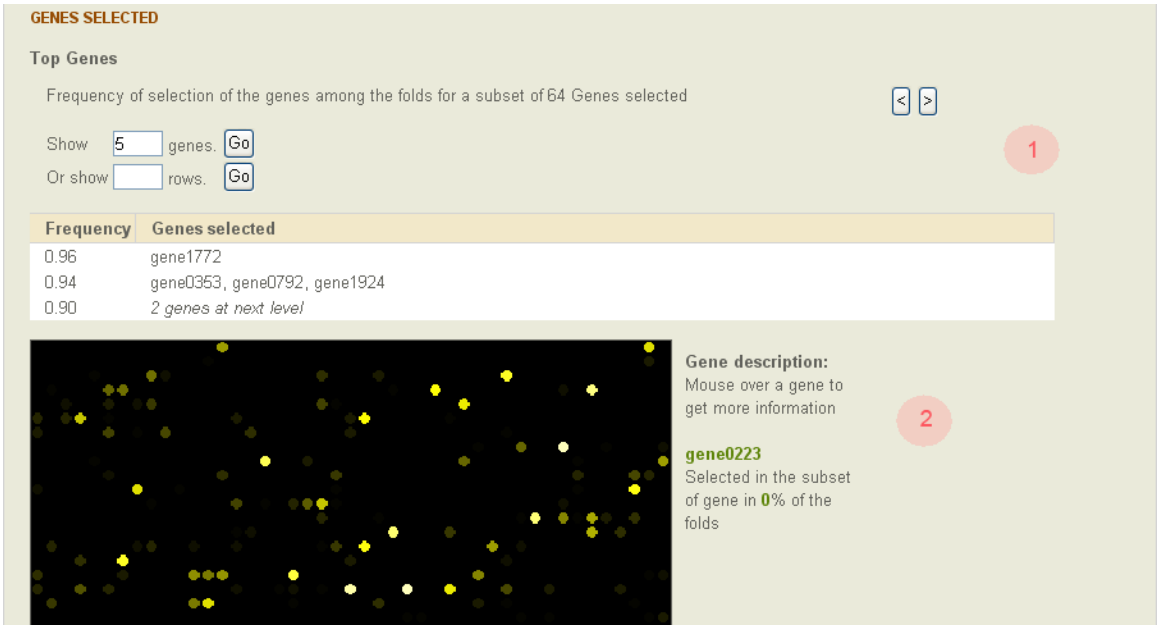



Figure 6.6: Summary view: Genes Selected

1. The user can navigate through the genes selected for different sizes of subsets or thresholds by clicking on these arrows.

2. 'Show X genes' allow the user to select the number of genes he wants to be displayed in the frequency table. By default a maximum of 100 genes are displayed. Figure 6.7 give an example where the user wants to see at most 20 genes.

3. 'show X rows' allow the user to select the number of rows he wants to be displayed in the frequency table. Figure 6.8 give an example where the user wants to see 3 rows only.

The last part of the 'Genes selected' section presents a visualization of the frequency of selection of each genes by the feature selection method. It is divided in two parts, highlighted in figure 6.9.

1. A microarray like image helps the user understanding the importance of each gene in the process of classification. The brighter is the dot the more influent is the corresponding gene.

2. A small description is available by mousing over a gene.

To generate the microarray like figure, the genes are ranked according to their frequency among the folds and the repeats. Then each gene is represented by a dot, the brightest dots correspond to the most frequent genes.


Figure 6.7: Top genes: Show the frequency of 20 genes only


Figure 6.8: Top genes: Show the frequency on 3 rows only

### Detailed view

As stated before, the detailed view can be accessed by selecting the 'Detailed view' link at the top-right corner of the 'Experimental results' section.

**Classification errors** The 'Classification errors' part is a bit more complicated as the one from the summary view. First, this part is divided in two subsections entitled 'Second layer of Cross-Validation' and 'First layer of Cross-Validation'.

**Second layer of cross-validation**

By default, cf. figure 6.10, the second layer of cross-validation is no further detailed, but the user can click on `+ More` to get the information concerning the performances on each fold of the outer layer.

Figure 6.11 present the expanded view of the second layer of CV and outlines four points of interest:

1. 'Less details' allows the user to get back to the compact second layer view
2. A graph plots the error rate and corresponding best number of genes or threshold for each fold and each repeat.
3. A table summarizes the error rate for each repeat and the corresponding average best number of genes or threshold.
4. A second table, at the bottom, displays the number and the names of the genes selected for each fold in each repeat.

Two tools, pointed out in figure 6.12, allow the user to interact with the genes table.

1. Arrow buttons to navigate through the repeats (similar as the one on the frequency table of the first Layer)


Figure 6.9: Visualization of the frequency of the genes

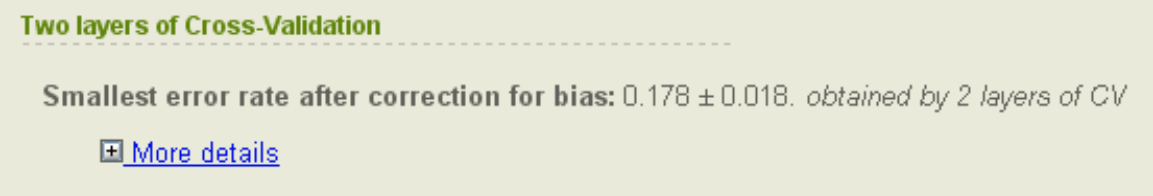



Figure 6.10: Classification Errors: Second layer of CV

2. By default, only 10 genes are displayed in each row, 'all genes' allow the user to have a look at all the genes selected for the current fold in the given repeat. Figure 6.13 gives an example where the first row has been expanded.

**First layer of cross-validation**

The First layer of cross-validation is very similar to the summary view. The differences are pointed out in figure 6.14.

1. The graph now plots the error rate versus the number of genes for each repeat of the first layer of cross-validation.

2. A radio button allows the user to switch views and display either the results of the aggregate (summary) first layer CV or of any of its repeats.

When the user decides to select the first repeat instead of the aggregate view, the error rate table changes to reflect the error rate of the first repeat only.

**Genes Selected** Figure 6.15 shows the display of the `Genes Selected` part in the detailed view. It's completely identical to the summary view excepted for a radio button at the top that allows the user to select which repeats (or aggregate) view he wants to display.

What happens when the user decides to display the genes selected in a repeat is shown in figure 6.16:

1. First, the frequency table is updated to show the genes selected during the first repeat only

2. Second, the microarray like image is removed, since it represents an overall summary of the implication of the genes in the outcome.

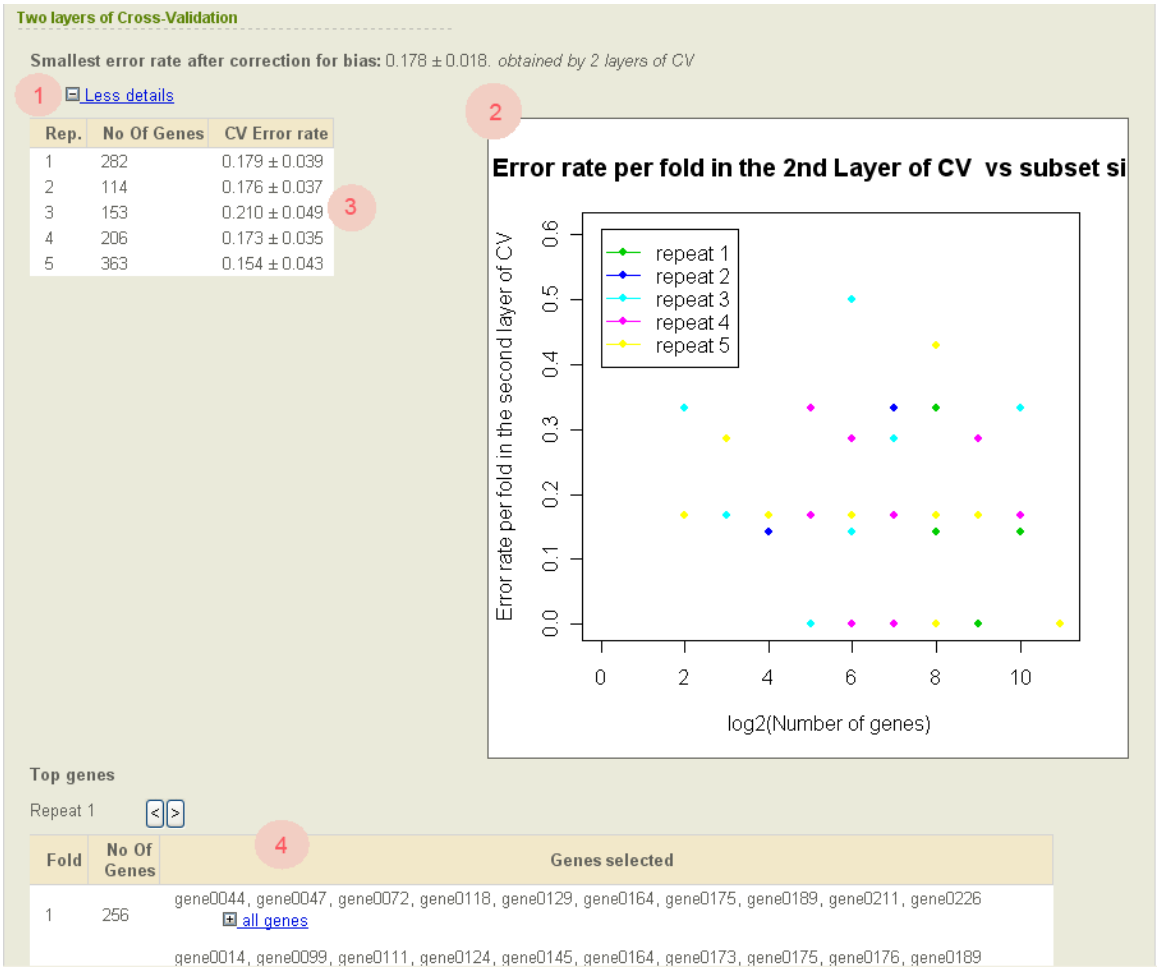



Figure 6.11: Second layer of CV: Expanded view

## 6.3 Delete an experiment

To delete an experiment click on the red cross corresponding to the one you want to delete from the list of experiments. As displayed in figure 6.17, a message wait for your confirmation before deleting the selected experiment. Click on `ok` to confirm the deletion or on `Cancel` if you want to go back the the list of experiments without deleting the selecting experiment.

## 6.4 Classify new samples

In order to classify new samples, you must first select an experiment which has already been performed. You can access to the page of classification of new samples by clicking on the link `Classify new samples` corresponding to the chosen experiment. Figure 6.18 displays the form that you have to fill to classify new samples. The view is divided in two parts:

1. The first frame described the options of the experiment selected
2. The second part contains the form

You must fill the form to classify one or more new samples by giving the following pieces of information:

- Give the path and the name of the file containing the levels of gene expression of the new samples that you want to classify


Figure 6.12: Second layer of CV: Genes selected

- Specify whether of not the names of the new samples are provided. If they are not, then they will be created automatically.
- Specify if you want to normalize (to a mean of zero and a standard deviation of one) the levels of gene expression of the new samples. You probably want to do the same treatment as the one you did on the original data.
- Specify whether each row corresponds to a gene or a sample.
- Finally, choose the size of subset or the threshold that you want to consider. If you don't know which one to choose, the best size of best threshold is probably a good trial.
- When you are ready click on the 'Classify' button

Figure 6.19 is an example of view of the results, after classification of new samples.

Repeat 3 < >

| Fold | No Of Genes | Genes selected |
|---|---|---|
| | | gene0014, gene0044, gene0124, gene0141, gene0142, gene0164, gene0175, gene0189, gene0211, gene0229 ⊟ 10 genes only |
| 1 | 128 | gene0249, gene0276, gene0280, gene0353, gene0377, gene0391, gene0419, gene0427, gene0448, gene0493, gene0510, gene0516, gene0533, gene0575, gene0576, gene0583, gene0611, gene0625, gene0642, gene0652, gene0654, gene0682, gene0698, gene0732, gene0739, gene0747, gene0752, gene0788, gene0792, gene0812, gene0822, gene0823, gene0835, gene0842, gene0890, gene0915, gene0925, gene0936, gene0955, gene0966, gene1005, gene1006, gene1013, gene1027, gene1030, gene1048, gene1060, gene1087, gene1094, gene1098, gene1110, gene1123, gene1147, gene1210, gene1221, gene1224, gene1226, gene1231, gene1241, gene1243, gene1256, gene1325, gene1346, gene1347, gene1348, gene1360, gene1366, gene1370, gene1400, gene1420, gene1423, gene1440, gene1465, gene1466, gene1482, gene1494, gene1501, gene1516, gene1530, gene1549, gene1550, gene1570, gene1579, gene1582, gene1597, gene1622, gene1625, gene1641, gene1668, gene1671, gene1679, gene1726, gene1727, gene1740, gene1756, gene1757, gene1769, gene1772, gene1780, gene1818, gene1823, gene1836, gene1843, gene1870, gene1873, gene1875, gene1892, gene1893, gene1895, gene1897, gene1909, gene1916, gene1920, gene1924, gene1935, gene1938, gene1983, gene1993 |
| 2 | 1024 | gene0003, gene0006, gene0011, gene0014, gene0015, gene0016, gene0025, gene0029, gene0030, gene0032 ⊞ all genes |
| 3 | 64 | gene0014, gene0099, gene0124, gene0164, gene0175, gene0202, gene0211, gene0249, gene0350, gene0353 ⊞ all genes |

Figure 6.13: Second layer of CV: Genes selected expanded

One layer of Cross-Validation

**Smallest error rate:** 0.152 ± 0.020 with **64** genes. *obtained by 1 external layer of CV*

Display Error Rates:
- (•) Summary
- ( ) Repeat 1
- ( ) Repeat 2
- ( ) Repeat 3
- ( ) Repeat 4
- ( ) Repeat 5

2

| No. Of Genes | CV Error rate | Class 1 | Class 2 |
|---|---|---|---|
| 1 | 0.255 ± 0.025 | 0.409 | 0.170 |
| 2 | 0.255 ± 0.025 | 0.409 | 0.170 |
| 4 | 0.223 ± 0.027 | 0.309 | 0.175 |
| 8 | 0.210 ± 0.017 | 0.282 | 0.170 |
| 16 | 0.219 ± 0.019 | 0.318 | 0.165 |
| 32 | 0.165 ± 0.023 | 0.227 | 0.130 |
| 64 | 0.152 ± 0.020 | 0.209 | 0.120 |
| 128 | 0.165 ± 0.019 | 0.218 | 0.135 |
| 256 | 0.152 ± 0.021 | 0.182 | 0.135 |
| 512 | 0.152 ± 0.019 | 0.200 | 0.125 |
| 1024 | 0.155 ± 0.021 | 0.200 | 0.130 |
| 2000 | 0.155 ± 0.023 | 0.191 | 0.135 |

Class Error Rates:
- [x] Class 1: Tumor
- [x] Class 2: Normal

1

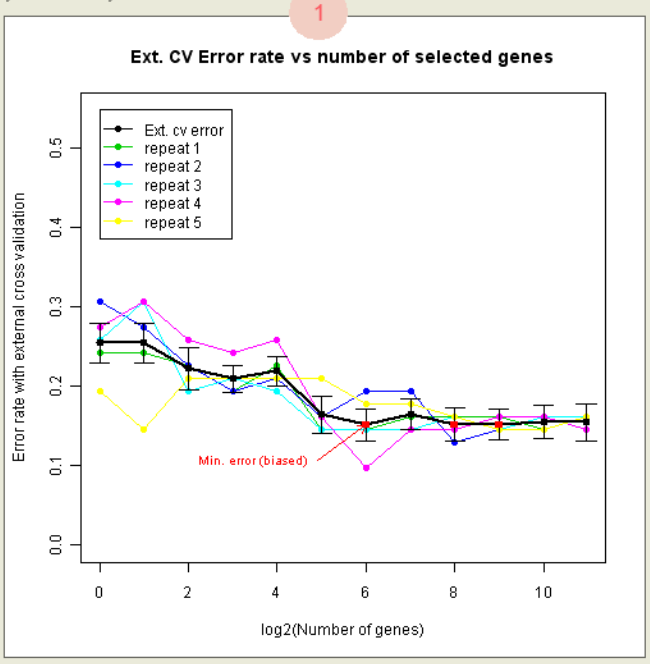



Figure 6.14: Classification errors: First layer of Cross-Validation

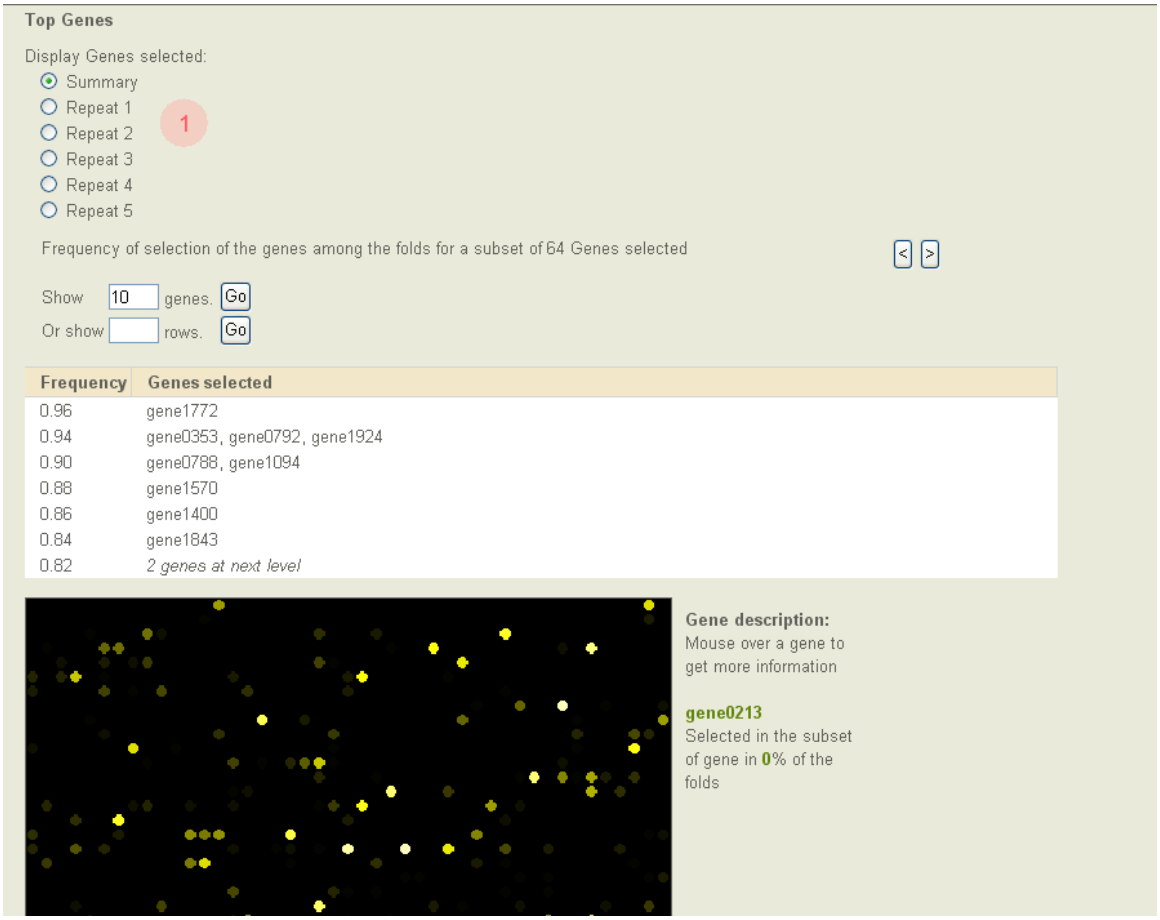



Figure 6.15: Detailed view: Genes selected


Figure 6.16: Detailed view: Genes selected for the first repeat

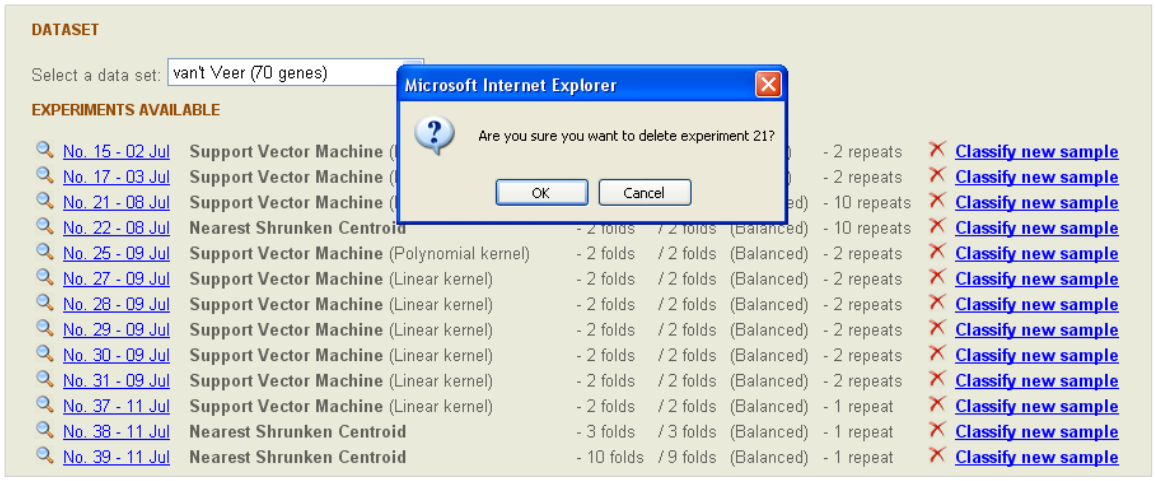



Figure 6.17: Deletion of an experiment


Figure 6.18: Form to classify new samples

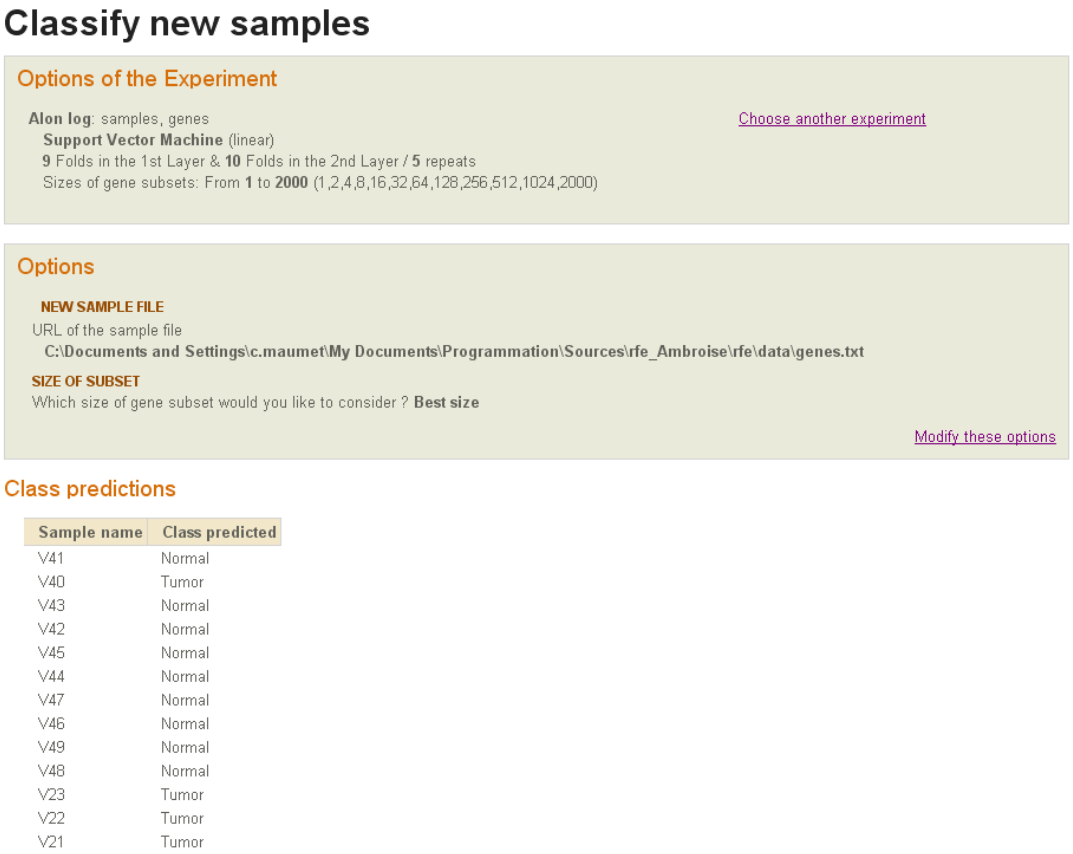



Figure 6.19: Result of the classification of new samples

# Chapter 7

# Conclusions

This document gave a detailed description of the Carrak software. If you have any comment or need more information please do not hesitate to contact us at Rmagpie@gmail.com.

# Bibliography

[Ambroise and McLachlan, 2002] Ambroise, C. and McLachlan, G. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences of the United States of America*, 99(10):6567–6572.

[Burman, 1989] Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514.

[Guyon et al., 2002] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422.

[Stone, 1974] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser.*, B(36):111–147.

[Tibshirani et al., 2002] Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 99(10):6567–6572.

[Wood et al., 2007] Wood, I., Visscher, P., and Mengersen, K. (2007). Classification based upon gene expression data: bias and precision of error rates. *Bioinformatics*, 23(11):1363–1370.

[Zhu et al., 2008] Zhu, J., McLachlan, G., Ben-Tovim, L., and Wood, I. (2008). On selection biases with prediction rules formed from gene expression data. *Journal of Statistical Planning and Inference*, 38:374–386.