ClaNC: The Manual (v1.1)

Alan R. Dabney[*†]

Department of Statistics, Texas A&M University, College Station, TX

[*]Department of Statistics, Texas A&M University, College Station, TX
[†]To whom correspondence should be sent: `adabney@stat.tamu.edu`

June 23, 2008

Contents

1   Installation

1.1   The R programming language

The ClaNC software is based in the freely-distributed R statistical computing language. Downloads and installation instructions for R can be found at:

$$http://www.r-project.org/$$

Click on CRAN under Download, then choose the appropriate pre-compiled version. Linux installation instructions are straightforward. Windows users should click on `Windows (95 or later)`, then click on `base/`, and download the set up executable. Mac OS X requires the following steps:

1. Download the latest R version from the `Mac OS X` link.

2. Double click on `R for Mac OS X 10.X.mkpg`, which will begin the installation routine.

3. When you get to the step `Installation Type`, click on `Customize`.

4. If `X11/TclTk (Jaguar)` is not selected, click on its box to select it.

5. Follow the rest of the instructions.

1.2   X11 with Mac OS X

Macintosh users will also need to download and install X11. X11 is used in the display of graphics for the ClaNC GUI and can be freely obtained from the Apple website:

$$http://www.apple.com/macosx/features/x11/$$

Click on `X11 for Mac OS X` in the right hand column. Fill out the registration information on the bottom of the page, and click `Download X11`. Start the installation software and follow the intuitive instructions.

1.3   The ClaNC package

The ClaNC package can be downloaded at

$$http://www.stat.tamu.edu/\sim adabney/clanc$$

3

Once R has been installed, the ClaNC package can be installed as follows:

- Windows: Start R and select the pull-down menu "Packages → Install package from local zipped file...", and then select clanc_X.Y.zip (with "X.Y" replaced by the appropriate version number).

- Mac / Unix / Linux: Type "R CMD INSTALL clanc_X.Y.tar.gz" (with "X.Y" replaced by the appropriate version number) at the command prompt in the directory where the file has been saved. For more information on INSTALL, start R and type "? INSTALL" at the R prompt.

2   Data Input

The following format for data input must be followed.

- Input file should be tab delimited text.

- All entries should be surrounded by quotes.

- Genes go in rows, and samples go in columns.

- First row contains class names for each sample (for training and test data).

- First column contains gene names.

- No missing values are allowed.

Example datasets can be found in the Data folder of the ClaNC directory after it has been installed. Note that this directory will typically be located in R's Library directory. If you have difficulty finding these files, they can also be downloaded from:

http://www.stat.tamu.edu/~adabney/clanc

3   Starting a ClaNC Session

First open R. If you are a Macintosh user, start X11. Then type the following commands: library(clanc) and clanc(). This will launch the ClaNC GUI, as in Figure 1.
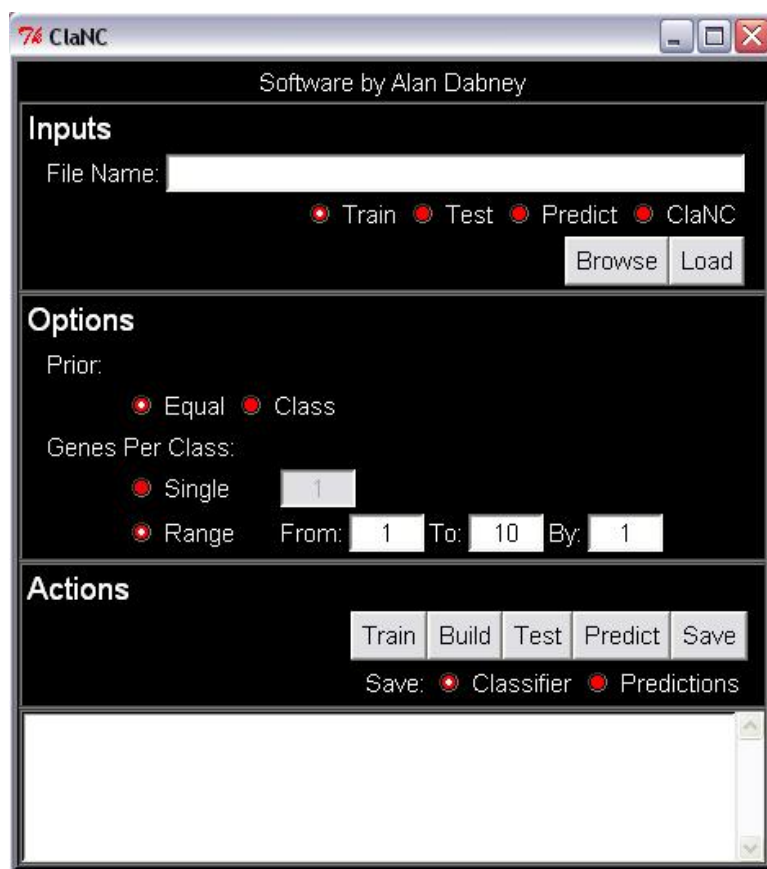
Figure 1: The ClaNC GUI.

4 Loading Data

You can load data sets for training, testing, or predicting. You can also load a previously saved ClaNC classifier. To load a file, select the file type, then click `Browse`. This will open a window for locating your file. Once you have selected the file, click `Load`. Depending on the size of the data set, this may take a while. A message will be posted in the box at the bottom when the action has completed. Please be patient.

5 Setting Options

You can choose two forms for the prior class probabilities. `Equal` will place equal weights on each class, while `Class` will choose weights that reflect the relative proportion of training samples within each class. You choose the number of active genes *in each class* to be either a single number or a

range of numbers. Typically, you will specify a range when training, then choose an optimal single number from the plot of cross-validated error rates when building the final classifier.

6   Training the ClaNC Classifier

The first step in building a classifier is training. With training data loaded, prior probabilities chosen, and active genes specified, click `Train`. This will compute some initial quantities and carry out five-fold cross validation. Again, depending on the size of the data set, this may take a while. If you specified a range of active genes, the cross-validated errors will be summarized by a plot, as in Figure 2. The plot can be resized by grabbing its lower right corner. It can also be saved by selecting `File`, then `Save` in R.
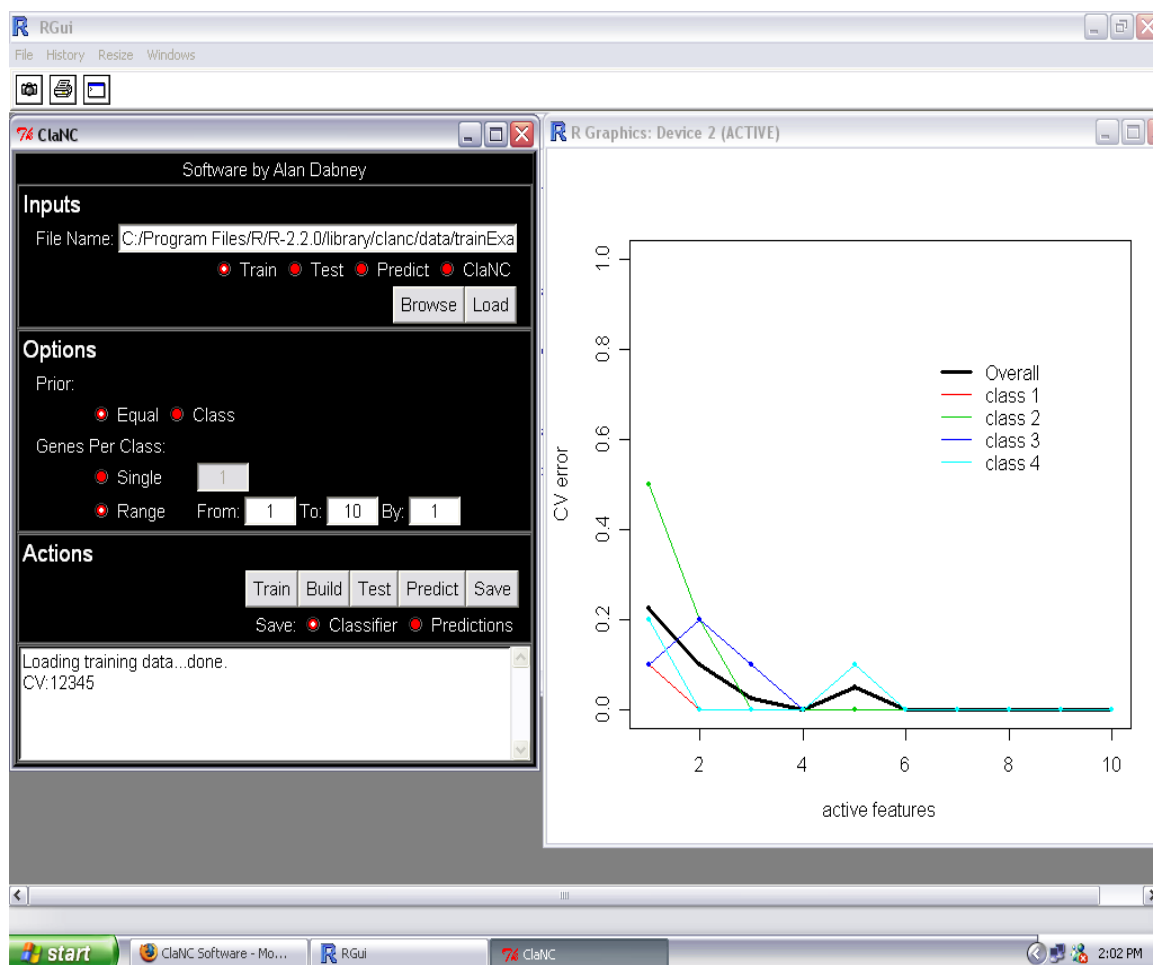

Figure 2: Cross-validated errors displayed after training on a range of active genes per class.

6

7 Building the ClaNC Classifier

Based on the cross-validated error rates, choose a single number of active genes per class, and update the options frame accordingly. You are now ready to build a ClaNC classifier. To do this, click `Build`. A plot will be formed to summarize the performance of your classifier on the training data, as in Figure 3. The training samples are grouped by their class membership along the x-axis, with the different classes separated by vertical lines. Correctly-classified samples have green points next to the correct class on the y-axis. Incorrectly-classified samples have red points next to the class to which they were assigned. The size of the classifier and the overall error rate is reported in the plot title.

Figure 3: Training errors for classifier built with user-specified number of active genes per class.

The resulting classifier can be saved to a tab-delimited text file by clicking `Save`, with `Classifier` selected below the `Save` button. The saved file contains prior probabilities stacked on top of a

matrix with (1) gene names, (2) pooled standard deviations, and (3) class centroid estimates for each gene used in the classifier. You can read this file back into a future ClaNC session program later by pressing the `Load` button with `ClaNC` selected in the `Inputs` frame.

8   Testing the ClaNC Classifier

After building the classifier, you may wish to test it on data that were not used in the training and building process. With the test data loaded, click `Test`. Once the action is completed, a summary of the test errors will be printed in the message box. A plot will also be formed to summarize the performance of your classifier on the test data, analagous to that in Figure 3.

9   Classifying New Samples

You can use the ClaNC classifier to predict the class membership of new samples. With prediction data loaded, click `Predict`. Once the action is completed, a summary of the classifications will be printed in the message box. You can also save this information to a tab-delimited text file by clicking `Save`, with `Predictions` selected below the `Save` button.

10   Getting Help

The official help forum is the Google discussion group. Please post any questions there. Membership is free.

<div align="center">

http://groups.google.com/group/clanc

</div>

11   Citing ClaNC

When referencing ClaNC, please cite one of Dabney (2005, 2006).

References

Dabney, A. R. (2005). Classification of microarrays to nearest centroids, *Bioinformatics* **21**: 4148–4154.

Dabney, A. R. (2006). ClaNC: Point-and-click software for classifying microarrays to nearest centroids, *Bioinformatics* **22**: 122–123.