BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 1 2006, pages 122–123
doi:10.1093/bioinformatics/bti756

*Gene expression*

# ClaNC: point-and-click software for classifying microarrays to nearest centroids

Alan R. Dabney

Department of Biostatistics, University of Washington, Seattle 98195, USA

Received on September 20, 2005; revised on January 10, 2005; accepted on October 27, 2005
Advance Access publication November 2, 2005
Associate Editor: John Quackenbush

**ABSTRACT**

**Summary:** ClaNC (classification to nearest centroids) is a simple and an accurate method for classifying microarrays. This document introduces a point-and-click interface to the ClaNC methodology. The software is available as an R package.

**Availability:** ClaNC is freely available from http://students.washington.edu/adabney/clanc

**Contact:** adabney@u.washington.edu

**Supplementary information:** http://students.washington.edu/adabney/clanc/figue2.pdf

In Dabney (2005), I described the classification to nearest centroids (ClaNC) method of classifying microarrays. ClaNC is derived from a classical method called linear discriminant analysis (LDA). LDA-based classifiers are simple and have been shown to outperform many more complicated alternatives (Dudoit *et al*., 2002; Lee *et al*., 2005; Tibshirani *et al*., 2002). ClaNC customizes LDA to the microarray problem by incorporating a feature-selection step. Briefly, ClaNC uses regular *t*-statistics to rank genes by their ability to distinguish the classes, then applies a class-specific procedure for selecting the genes to include in the classifier.

The prediction analysis of microarrays (PAM) method of Tibshirani *et al*. (2002) is another example of an LDA-based classifier for microarrays. ClaNC error rates tend to be substantially less than their PAM counterparts, for a particular number of genes used in the classifier (Dabney, 2005). Furthermore, ClaNC error rates tend to be less variable than PAM error rates (Dabney, 2005).

ClaNC runs on top of `R` (R Development Core Team, 2005, http://www.R-project.org) and is available as an `R` package. Detailed installation instructions can be found at my website. Once invoked, the ClaNC interface will appear as in Figure 1. A typical analysis will proceed as follows. First, load data to be used in training the classifier; all data input takes place in the top frame, labeled `Inputs`. Second, specify the prior probabilities to be used and the range of active genes per class to consider in the middle frame, labeled `Options`. Choosing `Equal` priors will place prior weight $1/K$ on each class, where $K$ is the number of classes. Choosing `Class` priors will place weight $n_k/n$ on class $k$, where $n_k$ is the number of training samples in class $k$, $k = 1, 2, \ldots, K$, and $n$ is the total number of samples. The number of genes to use in each class is specified directly. By default, classifiers using $1, 2, \ldots, 10$ genes in each class will be evaluated. To be clear, the classifier using 1 gene per class will choose $K$ unique genes, one unique entry for each class. Third, train the classifier by clicking `Train` in the bottom frame. This will carry out 5-fold cross-validation to assign misclassification error rates to each classifier under consideration. Once completed, a plot will appear summarizing the results, as in Figure 2 (Supplementary material).

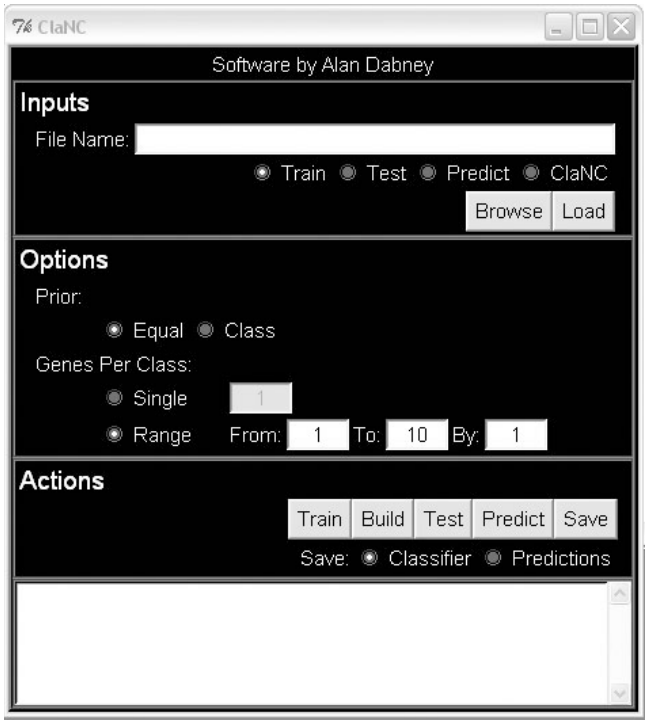


**Fig. 1.** The ClaNC interface.

Fourth, choose a single number of active genes per class based on this plot. In the example of Figure 2, four genes per class (16 total unique genes) drop the error rate to zero for all four classes. To build the classifier based on four genes per class, select `Single` under `Genes Per Class` in the `Options` frame, and click `Build`. Fifth, if you would like to evaluate your classifier on independent test data, load the data with `Test` selected in the `Inputs` frame, then click `Test`. A summary of the results will be printed in the message box. Sixth, if you would like to predict the class of an unknown sample (or a collection of unknown samples), load the data with `Predict` selected in the `Inputs` frame, then click `Predict`. The predicted classes will be printed to the message box. Finally, you can `Save` your classifier to a text file. Table 1 shows the results of saving the classifier built above. See the manual

 

**Table 1.** Example of a saved ClaNC classifier

| GeneNames | PooledSD | Class 1 | Class 2 | Class 3 | Class 4 |
|---|---|---|---|---|---|
| prior | NA | 0.25 | 0.25 | 0.25 | 0.25 |
| gene 363 | 1.230 | −0.495 | 1.003 | −0.495 | −0.495 |
| gene 502 | 1.069 | 4.172 | 1.743 | 1.743 | 1.743 |
| gene 751 | 1.179 | −0.173 | −0.173 | 1.444 | −0.173 |
| gene 756 | 1.061 | 1.850 | 1.850 | 0.402 | 1.850 |
| gene 823 | 0.852 | 0.942 | −1.203 | −1.203 | −1.203 |
| gene 833 | 1.142 | −2.027 | −0.354 | −2.027 | −2.027 |
| gene 894 | 1.290 | −1.555 | −1.555 | 0.353 | −1.555 |
| gene 907 | 1.425 | −0.271 | −0.271 | −0.271 | 1.698 |
| gene 959 | 1.322 | 0.939 | 0.939 | 0.939 | −0.679 |
| gene 1074 | 1.214 | 0.145 | 0.145 | 0.145 | 1.565 |
| gene 1077 | 1.337 | 0.831 | −0.859 | 0.831 | 0.831 |
| gene 1192 | 1.143 | −0.363 | −0.363 | 1.440 | −0.363 |
| gene 1258 | 1.256 | −1.253 | −1.253 | −1.253 | −3.059 |
| gene 1383 | 1.271 | −0.466 | −2.033 | −0.466 | −0.466 |
| gene 1412 | 0.855 | 2.170 | −0.187 | −0.187 | −0.187 |
| gene 1426 | 0.831 | −1.834 | −0.071 | −0.071 | −0.071 |

All components neccessary to represent the classifier are present: gene names, pooled standard deviations, the class centroid components for the selected genes, and the prior probabilities used.

for a more detailed discussion of the file format, but note that all components necessary to represent the classifier are present. This file can later be loaded into ClaNC by selecting `ClaNC` in the `Inputs` frame.

## ACKNOWLEDGEMENTS

This research was supported in part by the Cancer-Epidemiology and Biostatistics Training Grant 5T32CA009168-29 and NIH grant 1 U54 GM2119-03.

*Conflict of Interest:* none declared.

## REFERENCES

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

Dudoit,S. *et al.* (2002) Comparison of discriminant methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77–87.

Lee,J.W. *et al.* (2005) An extensive comparison of recent classification tools applied to microarray data. *Comput. Stat. Data Anal.*, **48**, 869–885.

R Development Core Team (2005) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Tibshirani,R. *et al.* (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.