A Short Introduction to the OmicsMarkeR Package

Charles Determan Jr.
cdetermanjr@hotmail.com

October 23, 2013

1 Introduction

The OmicsMarkeR package contains functions to streamline the analysis of 'omics' level datasets with the objective to classify groups and determine the most important features. OmicsMarkeR loads packages as needed and assumes that they are installed. I will provide a short tutorial using the both synthetic datasets created by internal functions as well as the 'Sonar' dataset. Install OmicMarkeR using

```
> # currently in development
> # install.packages("OmicsMarkeR", dependencies = c("Imports"))
```

to ensure that all the needed packages are installed.

You may also install the developmental version using

```
> library(devtools)
> install_github("OmicsMarkeR", username = "cdeterman")
```

2 Example

OmicsMarkeR has a few simplified functions that attempt to streamline the classification and feature selection process including the addition of stability metrics. We will first generate a synthetic dataset. This includes three functions that can be used to create multivariate datasets that can mimic specific omics examples. This can include a null dataset via `nvar = 50` and `nsamp = 100`.

The `create.corr.matrix` function induces correlations to the datasets. The `create.discr.matrix` function induces variables to be discriminate between groups. The number of groups can be specified with `num.groups`.

```
> library("OmicsMarkeR")
> set.seed(123)
> dat.discr <- create.discr.matrix(
+   create.corr.matrix(
+     create.random.matrix(nvar = 50,
+                          nsamp = 100,
+                          st.dev = 1,
+                          perturb = 0.2)))$discr.mat
```

To avoid confusion in the coding, one can isolate the variables and classes from the newly created synthetic dataset. These two objects are then used in the `fs.stability` function. I can then choose which algorithm(s) to apply e.g. `method = c("plsda", "rf")`, the number of top important features `f = 20`, the number of bootstrap reptititions for stability analysis `k = 3`, the number of k-fold cross-validations `k.folds = 10` and if I would like to see the progress output `verbose = TRUE`.

```
> vars <- dat.discr[,1:(ncol(dat.discr)-1)]
> groups <- dat.discr[,ncol(dat.discr)]
> results <- fs.stability(vars,
+                         groups,
+                         method = c("plsda", "rf"),
+                         f = 20,
+                         k = 3,
+                         k.folds = 10,
+                         verbose = FALSE)

[1] "plsda complete"
[1] "rf complete"
Aggregating results
Selecting tuning parameters
[1] "plsda complete"
Aggregating results
Selecting tuning parameters
[1] "rf complete"
Aggregating results
Selecting tuning parameters

Calculating Model Performance Statistics
```

If I would like to see the performance metrics, I can simply use the `performance.metrics` function. This will provide a concise data.frame of confusion matrix and ROC statistics.

```
> performance.metrics(results)

Model Performance Statistics

                       plsda         rf
Accuracy           1.00000000 0.93333333
Kappa              1.00000000 0.86111111
ROC.AUC            0.68959436 0.94444444
Sensitivity        1.00000000 0.94444444
Specificity        1.00000000 0.91666667
Pos Pred Value     1.00000000 0.94444444
Neg Pred Value     1.00000000 0.91666667
Accuracy SD        0.08164966 0.08164966
Kappa SD           0.17010345 0.17010345
ROC.AUC SD         0.17233103 0.17233103
Sensitivity SD     0.06804138 0.06804138
Specificity SD     0.10206207 0.10206207
Pos Pred Value SD 0.06804138 0.06804138
Neg Pred Value SD 0.10206207 0.10206207
```