10-22-2005

Dear GEMS users,

Discussions with and feedback from GEMS users revealed to us that some users use the system for analysis of extremely small sample datasets. A good heuristic rule for defining extremely small sample in the context of cancer microarray diagnosis/outcome prediction, is having <40-50 samples in the most prevalent classification category **or** having <N samples (where N is the number of cross-validation folds) in the most rare classification category.

We would like to offer a related note of caution: All cross-validation (as well as all known unbiased) classification error estimators may have high variance in such extremely small sample settings. This also applies to independent prospectively collected sample validation since the variance of the error in the independent set can also be very high. Notice that these limitations are not limitations of GEMS but universally apply to all such analyses (automatic or by hand) because of the laws that govern statistical sampling and estimation.

What this means practically?

- First, one should treat produced performance estimates with caution if the dataset has extremely sample size. In such situations findings should be interpreted as tentative and exploratory rather than conclusive.

- Second, when one sees large differences in performance estimates depending on cross-validation data split, this is a natural consequence of the small sample (and not a malfunction of the program).

- Third, you may be able to collapse (i.e., join together) classification categories and overcome the problem if in your experimental context collapsed categories make sense from a biological perspective.

- Most importantly, we recommend that you collect as much sample as possible given your experimental resource constraints.

Some insight on the limitations of small-sample analysis can be obtained from the following study and references therein: *Braga-Neto UM, Dougherty ER. Is cross-validation valid for small-sample microarray classification? Bioinformatics 2004; 20(3):374–80.*

With best regards and wishes for your research,

From the GEMS development team

Constantin Aliferis M.D., Ph.D.,
Assistant Professor of Biomedical Informatics and Cancer Biology,
Director, Discovery Systems Laboratory,
Department of Biomedical Informatics,
Vanderbilt University